

David E. Rowe
Tilman Sauer
Scott A. Walter
Editors

Beyond Einstein

Perspectives on Geometry,
Gravitation, and Cosmology in the
Twentieth Century

Einstein Studies

Editors: Don Howard Diana L. Kormos-Buchwald

Volume 14

- Volume 1:* Einstein and the History of General Relativity
Don Howard and John Stachel, editors
- Volume 2:* Conceptual Problems of Quantum Gravity
Abhay Ashtekar and John Stachel, editors
- Volume 3:* Studies in the History of General Relativity
Jean Eisenstaedt and A.J. Kox, editors
- Volume 4:* Recent Advances in General Relativity
Allen I. Janis and John R. Porter, editors
- Volume 5:* The Attraction of Gravitation: New Studies in the History of General Relativity
John Earman, Michel Janssen, and John D. Norton, editors
- Volume 6:* Mach's Principle: From Newton's Bucket to Quantum Gravity
Julian B. Barbour and Herbert Pfister, editors
- Volume 7:* The Expanding Worlds of General Relativity
Hubert Goenner, Jürgen Renn, Jim Ritter, and Tilman Sauer, editors
- Volume 8:* Einstein: The Formative Years, 1879–1909
Don Howard and John Stachel, editors
- Volume 9:* Einstein from 'B' to 'Z'
John Stachel
- Volume 10:* Einstein Studies in Russia
Yuri Balashov and Vladimir Vizgin, editors
- Volume 11:* The Universe of General Relativity
A.J. Kox and Jean Eisenstaedt, editors
- Volume 12:* Einstein and the Changing Worldviews of Physics
Christoph Lehner, Jürgen Renn, and Matthias Schemmel, editors
- Volume 13:* Towards a Theory of Spacetime Theories
Dennis Lehmkuhl, Gregor Schiemann, Erhard Scholz, editors
- Volume 14:* Beyond Einstein
David E. Rowe, Tilman Sauer, Scott A. Walter, editors

More information about this series at <http://www.springer.com/series/4890>

David E. Rowe • Tilman Sauer • Scott A. Walter
Editors

Beyond Einstein

Perspectives on Geometry, Gravitation,
and Cosmology in the Twentieth Century

Editors

David E. Rowe
Institut für Mathematik
Johannes Gutenberg-Universität
Mainz, Germany

Tilman Sauer
Institut für Mathematik
Johannes Gutenberg-Universität
Mainz, Germany

Scott A. Walter
Centre François Viète
Université de Nantes
Nantes Cedex, France

ISSN 2381-5833

Einstein Studies

ISBN 978-1-4939-7706-2

<https://doi.org/10.1007/978-1-4939-7708-6>

ISSN 2381-5841 (electronic)

ISBN 978-1-4939-7708-6 (eBook)

Library of Congress Control Number: 2018944372

Mathematics Subject Classification (2010): 01A60, 81T20, 83C47, 83D05

© Springer Science+Business Media, LLC, part of Springer Nature 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This book is published under the imprint Birkhäuser, www.birkhauser-science.com by the registered company Springer Science+Business Media, LLC part of Springer Nature.

The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

Einstein Studies Series Preface

Einstein Studies was launched in 1989 under the joint editorship of Don Howard and John Stachel, the founding editor of *The Collected Papers of Albert Einstein* and the Director of Boston University's Center for Einstein Studies, which served as the administrative home for the Einstein Studies series. The series was envisioned as a companion to the Einstein Papers Project, then also housed at Boston University, as a venue for the publication of scholarship relating to all aspects of the life and work of Albert Einstein, and as a tool for engendering and supporting an expanding community of scholars, especially younger scholars, working on such topics. Einstein Studies also aimed to be broadly interdisciplinary, featuring not only work on the history of science and technical work in physics, but also philosophy of science and social science perspectives on physics and its cultural embedding.

At the time of John Stachel's resignation as co-editor in 2017, the Einstein Studies series had published a total of thirteen volumes on topics ranging from Einstein and the History of General Relativity and Einstein's Formative Years to Mach's Principle and Einstein Studies in Russia. Included among those thirteen volumes is the rich and important volume one of Stachel's own collected papers, *Einstein from 'B' to 'Z.'* It would not be immodest to say that the series has realized its early ambition by helping to build what is now a thriving international scholarship focused on Einstein, a body of scholarship that is exemplary in its technical sophistication, historical depth, philosophical acuity, and cultural contextualization.

With Diana Kormos Buchwald's assumption of the role of co-editor, the vital connection between the Einstein Studies series and the Einstein Papers Project, which she directs, is reaffirmed. With the addition of a distinguished, international, editorial advisory board, the series is poised to play an even more prominent role in fostering the further expansion and enhancement of Einstein scholarship, with, again, special attention to nurturing each new generation of younger scholars as they enter the field. We eagerly invite proposals covering every part of the subject terrain.

Notre Dame, IN
Pasadena, CA

Don Howard
Diana Kormos Buchwald

Preface

The 17 papers gathered in this volume resulted from the conference “Beyond Einstein: Historical Perspectives on Geometry, Gravitation, and Cosmology in the Twentieth Century,” which took place during the week of 22–26 September 2008 at the Johannes Gutenberg-Universität in Mainz. This international conference brought together leading experts in physics, mathematics, cosmology, and the history and philosophy of these disciplines to address various developments that grew out of Einstein’s theory of relativity. Given the ambitious agenda, the speakers that week rose to the occasion and presented highly accessible talks to a large and generally appreciative audience (the original program is reproduced as an appendix). As the papers in this volume attest, the Mainz conference was a memorable, highly interdisciplinary event. Still, in some ways it resembled other conferences on the history of general relativity held over the last 30 years. For that reason, it is most appropriate that these proceedings should appear in the series *Einstein Studies*, which documents several other conferences on the history of relativity theory and its impact on modern physics.

A decade has now passed since these papers were originally presented in the form of informal talks, and so it should come as no surprise that several of the original submissions had to be updated or modified in order to take more recent developments into account. Indeed, two long-anticipated experimental findings from the last few years had to be addressed by a few of the authors for this volume, as these carry strong implications for future physical research. The first of these came in July 2012 when physicists at CERN’s LHC announced the detection of a particle similar to the Higgs boson predicted by the Standard Model. The second sensational finding—the direct observation of gravitational waves—was announced almost exactly 100 years after Einstein’s original paper that predicted this phenomenon on the basis of his general theory of relativity.

In arranging these papers, we have chosen to group them into five different sections with the following themes: Part I—Mathematical and Physical Underpinnings of Space-Time, Part II—Testing General Relativity and Rival Theories, Part III—Geometry and Cosmology, Past and Present, Part IV—Mathematical Motifs in General Relativity and Beyond, Part V—Quantum Gravity, Conformal Boundaries,

and String Theory. These categories give only a very rough indication of the actual topics discussed; however, and in several cases, readers will find related themes in papers we have placed in different sections. A few brief remarks about each of the contributions should thus be made in order to convey the overall flavor of the volume.

The three essays in Part I all deal with foundational issues connected with space-time. Scott Walter's essay, "Figures of Light in the Early History of Relativity (1905–1914)," deals with the immediate prehistory of general relativity by focusing on the various mathematical representations of light propagation during this period. He begins with Einstein's "Zur Elektrodynamik bewegter Körper" (1905), in which Einstein postulated the universal velocity of light for observers in inertial frames. Walter discusses this aspect of special relativity in terms of the form-invariance of light spheres, contrasting this with Poincaré's approach based on a "light ellipsoid." These constructs were soon overshadowed by a third figure of light, Minkowski's "lightcone," the fundamental object of space-time geometry. Several other investigators soon followed, as lightcones became a familiar tool in accounting for the relativity of simultaneous events, the properties of four-vectors, or the conformal features of space-time geometry. Nevertheless, all three optical structures continued to play their respective roles as Einstein's relativity principle gradually gained adherents. Walter notes how Ludwig Silberstein's textbook on relativity from 1914 makes use of all three figures of light, marking a convenient point of closure for this study, which includes ample references to recent historical work on the emergence and reception of Einstein's ideas.

Whereas the transmission of light signals formed one cornerstone in Einstein's famous 1905 paper, another came with his new kinematical principles based on measurements made with rods and clocks in inertial systems. When passing to general relativity, it has been customary to treat these as universal features of the metric field in a space-time manifold. Harvey Brown notes at the outset of his essay that experimental evidence seems to support the independence of measurements made by rods and clocks from their physical constitution. Nevertheless, he invites us to reconsider this aspect of Einstein's theory, starting with an examination of Einstein's own mature reflections on this question. After citing various experimental evidence for the deeper underpinnings behind the universal behavior of rods and clocks, Brown points to an oft-dismissed intellectual tradition that sought to explain contraction effects dynamically. In the 1970s, John S. Bell promoted this unorthodox approach, but Brown notes that Bell's ideas can be traced back to British ether theorists. Thus, Oliver Heaviside described the distortion of the electrostatic field that takes place when a point charge undergoes motion. This result was the inspiration behind G.F. FitzGerald's ideas about the deformation of rigid bodies as they move through the luminiferous ether. Taking this line of argument a step further, Brown writes: "For Einstein, Minkowski had done for relativity what Heaviside and others did for Maxwell theory when they introduced the three-vector formulation of electrodynamics (so that the physics is manifestly Euclidean covariant)." In stressing this rather than Minkowski's fusion of space and time, he sees Einstein as "distancing himself from his formulation of 1905 with its emphasis on fundamental

phenomenological postulates.” As applied to the metric field in general relativity, Brown contends that it may well be misleading to view $g_{\mu\nu}$ as a property inherent in space-time itself as opposed to a field over space-time.

In their essay “Hilbert on General Covariance and Causality,” Katherine Brading and Thomas Ryckman take up one of the themes in debates surrounding the mutual roles of Hilbert and Einstein during the period when general relativity was launched. In Einstein’s case, the critical turn took place in the fall of 1915, when he abandoned the Einstein-Grossmann equations and step-by-step found his way to the generally covariant equations familiar today. After his successful breakthrough to general covariance in November 1915, his infamous “hole argument” was resolved by the “point coincidence argument.” Hilbert had learned about Einstein’s theory and its problems from the latter’s lectures, delivered in Göttingen in the summer of 1915, and during the ensuing months he struggled to find his own way to come to a satisfactory resolution. As Brading and Ryckman point out, Einstein had two different arguments in support of the claim of restricted covariance, one based on energy conservation, the other his infamous “hole argument,” which purported to undermine causality. But they contend that the conceptual mistake of the latter, which was first analyzed in depth by John Stachel, was never a problem for Hilbert. The thrust of their argument is that Hilbert’s whole aim from the very beginning was to frame a well-defined Cauchy problem, which he initially believed required restricting the Einstein equations (or his version of these based on a variational principle). Hilbert’s specific view of causality in terms of the Cauchy problem remains recognizable after the final establishment of general relativity. To cite from their concluding remarks: “Hilbert’s ‘causality problem’ is not that which Einstein expressed in his ‘hole argument’ and resolved with his ‘point coincidence argument’. Rather, Hilbert *begins* from the key premise of the ‘point coincidence argument’, that coordinates are arbitrary labels, and seeks to solve the mathematical and epistemological problems that then arise.”

In Part II the focus shifts to the interplay between theory and observation. In “Putting General Relativity to the Test: 20th Century Highlights and 21st Century Prospects,” Clifford Will summarizes the present state of empirical support for the general theory of relativity, beginning with a number of high precision null experiments that strongly confirm Einstein’s equivalence principle. Further confirmation of general relativity has come from experiments in the solar system as well as tests calculating the post-Newtonian limit of metric theories. Measurements from binary pulsars have been made over several decades to test the possibility of gravitational-wave damping (see Dan Kennefick’s essay). This work served as a prelude to the direct detection of gravitational waves in 2015 by LIGO. Will points to three main areas for future testing: (1) strong-field gravity in the neighborhood of black holes and neutron stars, sources located far from the weak-field regime of the solar system; (2) the strong-field, radiative regime, which has now been opened with the new technologies that led to the LIGO breakthrough; and (3) gravity at extreme scales, where cosmological and micro-physical phenomena, possibly based on theories with additional dimensions, have yet to be explored.

Whereas Will deals mainly with recent work on the detection of gravitational effects, Herbert Pfister takes up one of the oldest problems that motivated Einstein's approach to gravitation: the relativity of rotation. This problem was rooted in Mach's ideas about the origins of inertia, which led Einstein to consider the possibility of deriving inertial effects by rotating a hollow sphere with a fictive test particle in its interior. With the advent of general relativity, he urged Hans Thirring to tackle this problem anew. Pfister notes that the work of Thirring and Josef Lense required that their isolated rotating mass shells be embedded in an asymptotically flat space-time. This limitation was later overcome, however, so that these same types of dragging effects could also be deduced in rotationally disturbed Friedmann–Lemaître–Robertson–Walker (FLRW) cosmologies. On the basis of recent observational findings, he concludes that “relativity of rotation” is perfectly realized in our universe. Indeed, in some cases, “the relative angular velocity between ‘local inertial systems’ and the most distant galaxies and quasars is estimated to be below 10^{-9} of the earth's angular velocity.” Unfortunately, Herbert Pfister did not live to see this essay in print, though Markus King kindly proofread the original version and updated it. Pfister was co-organizer of the international conference on Mach's principle, held in Tübingen in 1993, which led to another volume in Einstein Studies, *Mach's Principle: From Newton's Bucket to Quantum Gravity*. He is remembered in an obituary by Jörg Frauendiener, Domenico Giulini, and Markus King: “Nachruf auf Herbert Pfister,” *Physik Journal* 15 (2016) Nr. 1, S. 49.

In “Relativistic Lighthouses: The Role of the Binary Pulsar in Proving the Existence of Gravitational Waves,” Daniel Kennefick reviews some of the controversies that surrounded work on gravitational waves, particularly in connection with the discovery by Joseph Taylor and Russell Hulse in 1974 of the binary pulsar PSR 1913+16. His account makes ample use of the oral interviews he conducted with Taylor, Hulse, and others working in this field. Subsequent observations showed that the binary pulsar's orbit was gradually contracting, which was taken to be strong evidence for the emission of energy in the form of gravitational waves. Taylor and Hulse were able to show that the rate of decay of the orbit agreed with Einstein's theory based on the controversial quadrupole formula. In describing the background to this story, Kennefick relates these events to larger issues in the “Science Wars” debates that divided defenders of objective standards for consensus building from critics of this view, who argued that human and social factors were of decisive importance in science. In this context, Allan Franklin, taking the former view, attacked the general position of Harry Collins, who undertook a detailed analysis of the community of researchers who worked on gravitational waves. This topic had gained considerable notoriety in 1969 when the experimental physicist Joe Weber announced that he had detected and measured gravitational energy. Kennefick's main interest concerns the theoreticians and their changing views on the quadrupole formula in the ongoing controversies. He weighs in only lightly on the larger issues debated by Franklin and Collins, taking the position that what is most valuable comes from looking carefully at all the issues and actors that enter into this story, which of course has a happy ending. Recently, Franklin and Collins published a

joint paper in which they found some common ground but also agreed on their disagreement, see their “Two Kinds of Case Studies and a New Agreement,” in *The Philosophy of Historical Case Studies*, ed. Raphael Scholl and Tilman Sauer (Springer 2016, pp. 95–122). For parallel developments on efforts to represent gravitational radiation mathematically, the reader should turn to Jörg Frauendiener’s paper in Part V.

Allan Franklin, as noted above, has had a longstanding interest in understanding consensus building in experimental physics. In “The Rise and Fall of the Fifth Force,” he takes a close look at an important case study that might have had profound consequences for gravitational theory. As is well known, Einstein drew strong support from the Eötvös experiment in claiming that gravitational and inertial mass were identical. This led to his equivalence principle, already formulated in 1907, but which eventually became a cornerstone for the theory of general relativity. Proponents of the fifth force, however, disputed these findings. Franklin sets the background for this story, one strand going back to Ephraim Fischbach’s interest in whether gravitational effects might explain the observed violation of CP symmetry in K^0 meson decays. Fischbach conjectured that an additional term in a modified gravitational potential could be related to a small energy dependence in the CP-violating parameters in K^0 meson decay. In 1986, he, Sam Aronson, and Carrick Talmadge announced this proposed modification of Newton’s Law of Universal Gravitation, which made a splash in the popular press. Soon thereafter, experiments were performed in hopes of detecting a difference between the measured value of G by tests performed above ground and in mineshafts. The results, however, were inconclusive, making this the kind of case study Franklin excels in studying. He takes us into the details of this fast-moving story to show how by 1990 a strong consensus was reached: the Fifth Force, as imagined by Fischbach and others, did not exist.

Part III presents four studies that explore themes in modern cosmology and astrophysics, some of which have received little attention till now. In “Cyclic Models of the Relativistic Universe: The Early History,” Helge Kragh surveys various speculations regarding cyclic cosmologies in the period ca. 1922–1960. Alexander Friedmann, the pioneer of dynamic relativistic cosmology, was the first to realize that Einstein’s equations could lead to an expanding universe followed by contraction. George Gamow later found the notion of a big bang followed by a big crunch quite compelling. Richard Tolman even considered the possibility that the universe underwent a large or even infinite number of cycles. Whereas the notion of a cyclic universe was compatible with the Lemaître model and big-bang cosmology, it obviously contradicted the steady-state theory. Kragh emphasizes how cosmology after World War II was still strongly influenced by metaphysical and religious ideas. Cyclic cosmologies offered support for an eternal yet dynamic universe without a beginning in time. He points to Herman Zanstra as a prime example of a cosmologist who found cyclic models attractive because they conformed to his philosophical views. On the other hand, Zanstra rejected the possibility of a cosmic negative pressure, an idea that only gained acceptance later with the advent of inflationary cosmology.

Chris Smeenk's essay, "Inflation and the Origins of Structure," describes the evolution of this new approach to relativistic cosmology going back to Alan Guth's ideas from the late 1970s. Guth's original motivation for inflation was to solve the magnetic monopole problem, a major stumbling block for grand unified theories. An important consequence of the inflationary scenarios, however, only emerged later. Smeenk emphasizes how "inflation provides a mechanism for generating small departures from uniformity needed to seed formation of subsequent structures." Inflationary theory has been shrouded in controversy for many years. Some critics have assailed it as an "unfalsifiable" theory, whereas proponents have often pointed to the fact that it predicts spatial flatness. Smeenk suggests that these debates overlook larger issues that concern whether or not inflation offers an empirically testable account of the formation of cosmic structures. He arrives at this view after surveying some important historical background. A key event for the wider visibility of inflationary proposals was the Nuffield workshop held in Cambridge in 1982. Smeenk discusses three different approaches that emerged from this meeting as well as the loose consensus that emerged afterward when cosmologists developed a wide variety of inflationary models. Rather than assuming that an inflaton field would necessarily resemble the Higgs, it became more common to treat it as a new fundamental field distinct from any scalar field in particle physics. Smeenk also compares inflation with a competing theory for cosmological structure formation based on topological defects. Both approaches aimed to understand whether defects formed in the early universe could produce appropriate seeds for structure formation. Smeenk notes that observations of temperature fluctuations in the cosmic microwave background radiation discriminated between these two approaches, and clearly favored inflation.

The third paper in Part III, by Norbert Straumann, takes a critical look at various modifications of general relativity. A prime motivation for modified theories of gravitation comes from wanting to circumvent a major problem with the standard "concordance model" (Λ CDM model) of relativistic cosmology, namely its need to posit a massive amount of dark energy to account for the accelerated expansion of the universe. As Straumann shows, however, the alternatives to the standard model all have problems of their own. Still, he regards these ideas as important theoretically, since they show how modifications of the theory of general relativity can be used to gauge changes in the expansion rate of the universe. Any such modification has to meet stringent empirical tests, however, based on established astronomical findings for the solar system. Furthermore, a viable alternative model should be compatible with the cosmological data supporting the Λ CDM model. Straumann mainly focuses on $f(R)$ -metric theories, which arise by generalizing from the Ricci scalar R to a nonlinear function $f(R)$. This leads to a large family of alternative theories, from which one hopes to find a particular $f(R)$ that meets the above restrictions and which also can explain the accelerated expansion and structure formation of the universe without requiring dark energy. After this approach was first proposed in 1970, it became an active field of research in the 1980s following work by A.A. Starobinsky. Straumann concludes his analysis by

remarking that for the foreseeable future “it will be difficult to discriminate between dark energy and modified gravity, but this remains a major goal for years to come. One can hope that this will eventually become possible with better data on the CMB background, weak gravitational lensing, and the growth of large scale structures.”

In 1918 Hermann Weyl introduced what he called purely infinitesimal geometry, a generalization of Riemannian geometry based on a conformal metric. His motivation came from field physics: Weyl was at the time convinced that this approach would lead to a unification of Einstein’s gravitational theory with Gustav Mie’s matter theory based on a generalization of Maxwell’s equations. As is well known, Weyl soon gave up on this idea, though by the end of the 1920s his gauge-theoretic approach was successfully adapted in relativistic quantum physics. No one, however, including Weyl himself, saw any fruitful way to implement conformal metrics in a fruitful way for space-time physics. Yet, as Erhard Scholz demonstrates in his rich survey article, “Weyl Geometry in Late Twentieth Century Physics,” this abandoned notion rose from the ashes in the 1970s, when it was taken up independently by three different groups of authors. The first group—Jürgen Ehlers, Felix Pirani, and Alfred Schild—took up Weyl geometry in the context of their foundational research, using it as a conceptual framework for clarifying the foundations of gravity. A second group of authors was interested in using its generalized geometrical structure to develop extended gravitational theories, and a third group, centered around Wolfgang Drechsler and Hanno Tann in Munich, was primarily interested in exploring connections between gravitation and quantum physics. Scholz offers a sweeping account of subsequent developments that relate to several areas of active research. Among the many research groups that enter this story, he points to the work of the Brazilian school, led by Mário Novello, who in 2003 became the founding director of the *Instituto de Cosmologia Relatividade e Astrofísica (ICRA)*. During the course of the 1990s, Novello introduced Weyl-geometric ideas to several theoretical physicists in the Brazilian community. In summing up, Scholz writes that “since the 1970s Weyl-geometric approaches to gravity, elementary particle fields, foundations of quantum mechanics, astrophysics and cosmology have developed a rich array of models. . . . Although the Weyl-geometric perspective has remained up to now a side-stream in all of these fields, it may well offer interesting challenges and possibilities for the future.”

Part IV addresses topics that involve an especially strong interplay between mathematical and physical ideas. In “Matter from Space,” Domenico Giulini offers an overview of geometrodynamics, a trend that goes back to William Kingdon Clifford, one of the nineteenth century’s foremost advocates of a geometrized physics in the spirit of Riemann. Giulini admits that there is no hope of implementing this vision completely, but he cites Einstein’s 1920 Leiden lecture (as well as a famous cartoon in *The New Yorker*) to argue that geometrodynamics represents a core conception that is highly compatible with the philosophical tenets of general relativity. As he puts this, “it also seems to be a straightforward consequence of modern field theory, according to which fundamental fields are directly associated with space (or space-time) rather than any space-filling material substance. Once the latter view

is adopted, there seems to be no good reason to neglect the field that describes the geometry of space.” As examples of the utility of these ideas, Giulini points to various ways properties of matter can be modeled using matter-free gravitational fields, techniques that have been employed in studying the scattering and merging processes of black holes. He notes further, how “the enormous topological variety and complexity of 3-manifolds leave their structural traces in General Relativity,” a prime example being studies of structures that arise with wormholes in space-times.

In speaking of 3-manifolds, it is well to keep in mind that even the very simplest case, the three-sphere S^3 , led to a truly amazing series of events before its characteristic properties, first investigated by Henri Poincaré, were finally proven. The essence of that story can be found in Donal O’ Shea’s article, “The Unexpected Resolution of the Poincaré Conjecture.” Throughout the twentieth century, the Poincaré Conjecture was the outstanding open problem in topology, a tantalizingly simple question that defied all efforts to prove it or find some strange counterexample. Like \mathbf{R}^3 , an S^3 is simply connected, so every closed loop in it can be continuously deformed to the initial point of the loop. But unlike \mathbf{R}^3 , it is a closed manifold, and so it was natural to ask: is every closed, simply connected 3-manifold homeomorphic to the three-sphere? The answer finally came in 2003, when Grigory Perelman posted three papers on arXiv.org in which he proved the Poincaré conjecture. It took another three years for experts in the field to pronounce on his complicated and novel proof of what is now known as the Poincaré theorem. What astounded the mathematical world, however, was not the answer but rather the way Perelman went about proving this purely topological result, for which he drew both on Thurston’s work on 3-manifolds and Hamilton’s studies of Ricci flows. Moreover, as O’Shea points out, Perelman clearly had a broader vision in mind, a mathematical model of the universe based on a hierarchy of manifolds at different scales connected by means of the Ricci flow. He concludes by remarking: “No one, least of all Poincaré, would have ever imagined that techniques from analysis and mathematical physics to which topology had contributed so much, would one century later repay the favor by being used to solve the most famous purely topological problem of all time. And no one would have predicted that the resulting gadget, a hierarchy of Riemannian manifolds connected by the Ricci flow, might provide a mathematical object useful for modeling space and space-time at different scales.”

Mathematical motifs, especially those involving geometrization, also played a major role in research on unified field theories. This is the theme of Hubert Goenner’s study, which looks at work done on efforts to find a unified theory of the gravitational and electromagnetic fields and the leading communities that pursued this research up to the 1960s. These developments thus precede those described in Scholz’s paper in Part III. Einstein’s name looms very large in this picture, of course, as the problem of uniting the electromagnetic and gravitational fields dominated his attention almost exclusively once he left the field of quantum physics to a younger generation. Among the many questions Goenner raises, one concerns the extent to which Einstein’s authority and fame lent this research a special

attraction. Other major figures who entered this arena were Erwin Schrödinger in 1943, and soon thereafter Marie-Antoinette Tonnelat and Vaclav Hlavatý. Goenner describes efforts to solve the weak field equations in what was called the “Einstein-Schrödinger theory.” Over time, though, these were modified in the face of negative physical results. In the closing section, he offers a brief overview of the worldwide community of researchers working on unified field theories.

Finally, Part V presents three essays on three important areas of research that round out the volume. Among the various approaches to quantum gravity, Robert Wald offers reflections on how quantum field theory can be adapted directly to curved space-times. In doing so, however, significant changes have to be made. He notes that the usual terminology in quantum field theory leaves the impression that it is fundamentally a theory of “particles.” Yet, even for a free field, differentiating between “vacuum” and “particles” depends on being able to decompose the field into its positive and negative frequency parts, which requires exploiting the time translation symmetry in Minkowski space-time. Since no such symmetry is available in a generic space-time, there is no natural notion of a “vacuum state” or of “particles.” Wald focuses on the Wightman axioms for quantum field theory in a Minkowski space, showing why this approach presents significant difficulties in space-times with curvature. Nevertheless, these can be overcome by means of appropriate conceptual generalizations, the only exception being Wightman’s axiom for the existence of a unique, Poincaré invariant vacuum state. Wald likens the quest for a “preferred vacuum state” in quantum field theory to similarly futile quests for a “preferred coordinate system” in classical general relativity. He remarks in conclusion: “the attempt to describe quantum field phenomena in curved space-time has directly led to a viewpoint where symmetries and notions of ‘vacuum’ and ‘particles’ play no fundamental role. The theory is formulated in a local and covariant manner in terms of the quantum fields. This formulation is very well suited to investigation of quantum field effects in the early universe.”

In “Conformal Infinity—Development and Applications,” Jörg Frauendiener takes up important mathematical developments connected with understanding the nature of gravitational radiation. His account describes how early work on gravitational waves served as the motivating problem that eventually led to a new mathematical construct: conformal infinity, introduced by Roger Penrose in the 1960s. Einstein and Rosen had famously concluded in 1936 that gravitational waves do not exist, but Ivor Robinson and Herman Bondi showed the flaw in their reasoning, which stemmed from an overly restrictive regularity condition that required the entire space-time manifold to be covered by a single coordinate chart. The situation was not unlike earlier misinterpretations of the Schwarzschild solution that arose from reading too much into a chosen coordinate system. Felix Pirani was the first who saw the need for an invariant-theoretic characterization of gravitational radiation, which he introduced by means of the Riemann tensor. In 1958, Andrzej Trautman used a special class of coordinate systems, in which the metric approaches the flat Minkowski metric asymptotically to produce a well-defined 4-vector for the energy-momentum along hypersurfaces. Frauendiener describes various other

fast-breaking progress in this field, citing papers by Ray Sachs, Ted Newman, and Penrose. He then briefly summarizes the basic concepts Penrose developed for attaching conformal boundaries to space-times, discussing issues surrounding the requirements for such space-times to exist, and ending with a brief account of applications of their use. In conclusion, Frauendiener writes: “Penrose’s notion of conformal infinity has brought a completely new insight into the geometry of the asymptotic regions of an asymptotically flat space-time. It has changed the way we look at such space-times today and it forms the foundation for many developments, such as Newman’s \mathfrak{H} -space, Penrose’s twistor theory, and Ashtekar’s asymptotic quantisation procedure. It arose in the attempt to rigorously understand the nature of gravitational waves.”

The final paper, “String Theory and Space-Time Geometry” by Matthias Gaberdiel, offers a brief look at this fast-moving field that in recent years has attracted hundreds of young mathematical physicists. Part of this attraction stems from the surprising new insights that string theory has brought to topics in mathematics, in particular, algebraic geometry (mirror symmetry), group theory (monstrous moonshine), and number theory (modular forms). But Gaberdiel also emphasizes that string theory offers a convincing explanation for black hole entropy in terms of microstates. He describes the general field of string theory as “a quantum theory in which the different elementary particles and their interactions arise from one common principle.” In particular, string theory leads to a quantum theory of gravity since it contains gravitons, which correspond to the fluctuations of the initial background geometry. Gaberdiel begins his survey with bosonic string theory before turning to superstrings and D-Branes. He sees string theory as a promising approach that could bring together the now firmly grounded standard model for interactions of elementary particles with Einstein’s general relativity.

The original inspiration for this meeting came from conversations that the first of us had with Erhard Scholz in Wuppertal. His longstanding interest in Hermann Weyl’s work had brought him into contact with more recent attempts (including his own) to adapt Weyl’s purely infinitesimal geometry to the new vistas in field physics and cosmology. The idea to stage a conference in Mainz also fell nicely in place since 2008 was celebrated throughout Germany as “das Jahr der Mathematik.” In Mainz, Volker Bach threw his support behind this project from the beginning; his talents as a mathematical physicist, but also as a fundraiser, were critical for its success. We recall with thanks the various sponsoring institutions whose financial support made the conference possible: the German Research Foundation (DFG), the Ministry of Education of the state of Rheinland-Pfalz, the Johannes Gutenberg-Universität Mainz, and finally its Institute of Mathematics and its affiliated Special Research Group (Sonderforschungsbereich).

It remains to thank some of the many other people who helped make this event possible, starting with Dan Kennefick and Hubert Goenner, who provided us with valuable input during the planning stage of the meeting. Two theoretical physicists from Mainz, Florian Scheck and Martin Reuter, provided valuable help, as did Duco van Straten, Stefan Müller-Stach, and Manfred Lehn. Martin Bauer offered

his expertise in preparing the original papers in LaTeX, and in the further production process we also received help from Renate Emerenziani and Natalia Poleacova. Indeed, a special *danke schön* goes to Renate Emerenziani, who went out of her way to manage all the little things that conference organizers often tend to forget.

Mainz, Germany
Mainz, Germany
Nancy Cedex, France

David E. Rowe
Tilman Sauer
Scott A. Walter

Contents

Part I Mathematical and Physical Underpinnings of Spacetime

1	Figures of Light in the Early History of Relativity (1905–1914)	3
	Scott A. Walter	
1.1	Introduction.....	3
1.2	Einstein’s Light Sphere	5
1.3	Poincaré and the Lorentz Group	9
1.4	Langevin’s Electron Wake	11
1.5	Poincaré’s Light Ellipse.....	12
1.6	Minkowski’s Lightcone	22
1.7	Alfred A. Robb: Repurposing the Lightcone	28
1.8	Applications of the Light Sphere	33
1.9	Light-Figure Skepticism	40
1.10	Discussion	42
	References.....	44
2	The Behaviour of Rods and Clocks in General Relativity and the Meaning of the Metric Field	51
	Harvey R. Brown	
2.1	Introduction.....	51
2.2	Clocks and Their Complications	52
2.3	Special Relativity.....	55
2.4	General Relativity	60
	References.....	64
3	Hilbert on General Covariance and Causality	67
	Katherine Brading and Thomas Ryckman	
3.1	Introduction.....	67
3.2	Einstein’s “Causality Problem”.....	68
3.3	Einstein’s “Point Coincidence Argument”	69
3.4	Hilbert’s “Causality Problem”, 1915	70
3.5	Hilbert’s “Causality Problem”, 1917	71

3.6	Hilbert's 1917 Resolution of His "Causality Problem"	72
3.7	Hilbert and Einstein Compared	74
3.8	Conclusion	75
	References	75
Part II Testing General Relativity and Rival Theories		
4	Putting General Relativity to the Test: Twentieth-Century Highlights and Twenty-First-Century Prospects	81
	Clifford M. Will	
4.1	Introduction	81
4.2	Twentieth-Century Highlights	83
4.3	Twenty-First-Century Prospects	92
4.4	Conclusions	94
	References	95
5	Rotating Hollow and Full Spheres: Einstein, Thirring, Lense, and Beyond	97
	Herbert Pfister	
5.1	Introduction	97
5.2	Einstein's 1913 Work on Rotating Spheres	99
5.3	The Contributions of Thirring and Lense from the Years 1917 to 1918 and Their Dependence on Einstein's Intervention ...	100
5.4	Deficiencies of the 1918 Papers by Thirring and Lense and Their Corrections	102
5.5	The Solution of the Centrifugal Force Problem	104
5.6	A Quasiglobal Principle of Equivalence	105
5.7	Cosmological Considerations	107
	References	109
6	Relativistic Lighthouses: The Role of the Binary Pulsar in Proving the Existence of Gravitational Waves	111
	Daniel Kennefick	
6.1	Introduction	111
6.2	Controversy	113
6.3	Discovery	117
6.4	Trading Zones and Pidgins	120
6.5	Skeptics' Dilemma	123
6.6	Theory Testing	129
6.7	Conclusions	132
	References	135
7	The Rise and Fall of the Fifth Force	137
	Allan Franklin	
7.1	The Rise	138
7.2	... and Fall	145

7.3 Epilogue: The Fifth Force Since 1991 157
 References 175

Part III Geometry and Cosmology, Past and Present

8 Cyclic Models of the Relativistic Universe: The Early History 183
 Helge Kragh
 8.1 Introduction 183
 8.2 The Friedmann-Einstein Universe 184
 8.3 A Controversial Universe 189
 8.4 Richard Tolman and Cosmic Entropy 192
 8.5 The Oscillating Universe in the 1950s 196
 8.6 Negative Pressure as a Saving Device 200
 References 201

9 Inflation and the Origins of Structure 205
 Chris Smeenk
 9.1 Introduction 205
 9.2 Structure Formation in the Standard Model 207
 9.3 Inflationary Cosmology 215
 9.4 Demise of a Rival Approach: Topological Defects 227
 9.5 Characterizing Empirical Success 233
 References 237

10 Problems with Modified Theories of Gravity, as Alternatives to Dark Energy 243
 Norbert Straumann
 10.1 Introduction 243
 10.2 Metric $f(R)$ Gravity 244
 10.3 Generalized Friedmann Models 247
 10.4 Weak-Field Limit for Spherically Symmetric Sources 248
 10.5 Chameleon Mechanism 249
 10.6 Nonexistence of Relativistic Stars in $f(R)$ Gravity? 253
 10.7 Inclusion of Other Curvature Invariants 255
 10.8 First-Order (Affine) Modifications of GR 255
 10.9 Concluding Remarks 257
 References 258

11 The Unexpected Resurgence of Weyl Geometry in late 20th-Century Physics 261
 Erhard Scholz
 11.1 Introduction 261
 11.2 Preliminaries: Weyl-Geometric Gravity and Jordan-Brans-Dicke Theory 264
 11.3 Contributions to Weyl-Geometric Gravity in the 1970s and 1980s 273

- 11.4 Weyl’s Scale Connection a Geometrical Clue to Quantum Mechanics? 291
- 11.5 Scale Covariance in the Standard Model of Elementary Particle Physics 307
- 11.6 Weyl-Geometric Models in Astrophysics and Cosmology Since the 1990s 328
- 11.7 Discussion 345
- References 350

Part IV Mathematical Motifs in General Relativity and Beyond

- 12 Matter from Space** 363
 - Domenico Giulini
 - 12.1 Introduction 363
 - 12.2 Geometrodynamics 365
 - 12.3 X Without X 382
 - 12.4 Further Developments 392
 - References 397
- 13 The Surprising Resolution of the Poincaré Conjecture** 401
 - Donal O’Shea
 - 13.1 Introduction 401
 - 13.2 Three-Dimensional Manifolds and the Poincaré Conjecture 402
 - 13.3 Poincaré’s Topological Papers 403
 - 13.4 Perelman’s Proof 409
 - 13.5 Conclusion 413
 - References 414
- 14 Unified Field Theory up to the 1960s: Its Development and Some Interactions Among Research Groups** 417
 - Hubert Goenner
 - 14.1 Introduction 417
 - 14.2 The Geometrization of Physics 419
 - 14.3 Field Equations in Unified Field Theory, and How to Solve Them 423
 - 14.4 Persons, Publications, Interactions 430
 - 14.5 Conclusion 433
 - References 434

Part V Quantum Gravity, Conformal Boundaries, and String Theory

- 15 The Formulation of Quantum Field Theory in Curved Spacetime** ... 439
 - Robert M. Wald
 - References 448

- 16 Conformal Infinity – Development and Applications** 451
 - Jörg Frauendiener
 - 16.1 Introduction 451
 - 16.2 Towards the Invariant Characterisation of Gravitational Waves ... 452
 - 16.3 Asymptotic Structure and Conformal Geometry 459
 - 16.4 Existence of Asymptotically Flat Space-Times 465
 - 16.5 Applications 468
 - 16.6 Conclusion 470
 - References 471

- 17 String Theory and Spacetime Geometry** 475
 - Matthias R. Gaberdiel
 - 17.1 Generalities 475
 - 17.2 Bosonic String Theory 478
 - 17.3 The Superstring 481
 - 17.4 D-Branes 485
 - 17.5 Conclusions 488
 - References 489

- Conference Program** 491
 - Volker Bach and David E. Rowe

Part I
Mathematical and Physical Underpinnings
of Spacetime

Chapter 1

Figures of Light in the Early History of Relativity (1905–1914)



Scott A. Walter

1.1 Introduction

When Albert Einstein first presented his theory of the electrodynamics of moving bodies (Einstein 1905), he began by explaining how his kinematic assumptions led to a certain coordinate transformation, soon to be known as the “Lorentz” transformation. Along the way, the young Einstein affirmed the form invariance of the equation of a spherical light wave (or light-sphere covariance, for short) with respect to inertial frames of reference. The introduction of the notion of a light sphere in this context turned out to be a stroke of genius, as Einstein’s idea resonated with physicists and mathematicians, and provided a way to understand the Lorentz transformation, kinematics, simultaneity, and Lorentz covariance of the laws of physics.

A focus on the light sphere as a heuristic device provides a new perspective on the reception of relativity theory and on the scientific community’s identification of Einstein as the theory’s principal architect. Acceptance of relativity theory, according to the best historical accounts, was not a simple function of having read Einstein’s paper on the subject.¹ A detailed understanding of the elements that turned Einsteinian relativity into a more viable alternative than its rivals is, however, not yet at hand. Likewise, historians have only recently begun to investigate how scientists came to recognize Einstein as the author of a distinctive approach to relativity, both from the point of view of participant histories (Staley 1998) and

¹For gradualist views of the acceptance of relativity theory see Hiosige (1968), Miller (1981), and Darrigol (1996, 2000).

S. A. Walter (✉)
Centre François Viète, Université de Nantes, Nantes Cedex, France
e-mail: scott.walter@univ-nantes.fr

from that of disciplinary history (Walter 1999a). The latter studies underline the need for careful analysis when evaluating the rise of Einstein's reputation in the scientific community, in that this ascent was accompanied by that of relativity theory itself.

We know, for example, that the fortunes of relativity theory improved when Bucherer (1908a) announced the results of electron-deflection experiments in line with relativist predictions. Einstein's most influential promoter, Max Planck, himself a founder of relativistic dynamics, was in Einstein's view largely responsible for the attention paid by physicists to relativity theory (Heilbron 1986, 28). Planck also praised Hermann Minkowski's four-dimensional approach to relativity, the introduction of which marked a turning point in the history of relativity (Walter 1999a). There is more than Planck's praise to tie Einstein's theory of relativity to Minkowski's spacetime theory. Much as the lightcone distinguishes Minkowski's theory from earlier theories of space and time, the light sphere was one of the key objects that set apart Einstein's theory of relativity (as it became known around 1911) from alternative theories of the electrodynamics of moving bodies.

My account begins with Einstein's relativity paper of 1905, in which the notion of the form invariance of the equation of a light sphere was introduced. While interest in form invariance of the differential equation of light-wave propagation dates from the 1880s, the idea that a light sphere remains a light sphere for all inertial observers – with a universal velocity of light – was recognized as a major conceptual innovation in the fall of 1907, when it was first used to derive the Lorentz transformation. By then, the light sphere had already been employed in Paris by Henri Poincaré, along with a second figure of light, the “light ellipsoid,” to illustrate an alternative to Einsteinian kinematics. Inspired by his readings of Einstein and Poincaré, Minkowski identified and exploited a third figure of light, the “lightcone,” to define and illustrate the structure of spacetime. In the wake of spacetime theory, other investigators used figures of light to explore the relation of simultaneity, the properties of four-vectors, and the conformal structure of spacetime. The period of study comes to a close with the publication of Ludwig Silberstein's textbook on relativity, which was the first to feature all three figures of light. Although light figures sparked discussion and debate until the early 1920s, Silberstein's discussion represents a point of closure on this topic, by bringing together previously disjoint intellectual developments of the previous decade.

By following light figures through a selection of published and archival sources during the period 1905–1914, the skills and concerns of a nascent community of relativists are brought into focus. The progress of this community's knowledge of the scope, history, and foundation of relativity theory, as it related to the domains of measurement theory, kinematics, and group theory, is reflected in the ways it put these new objects to use, by means of accounts both formal and discursive in nature. During the formative years of relativity, an informal, international, and largely independent group of physicists, mathematicians, and engineers, including Einstein, Paul Langevin, Poincaré, Minkowski, Ebenezer Cunningham, Harry Bateman, Otto Berg, Max Planck, Max von Laue, Arthur A. Robb, and Ludwig Silberstein,

employed figures of light to discover salient features of the relativistic worldview. Their contributions, and those of their critics, are considered here on their own merits, as part of an intellectual movement taking place during a period when the meaning of the theory of relativity was still negotiable, and still being negotiated.

1.2 Einstein's Light Sphere

The concepts of relative time and relative simultaneity were taken up by Einstein in the course of his relativity paper of 1905. It seems he was then unaware of Lorentz's (1904) attempt to demonstrate the form invariance of Maxwell's equations with respect to the Lorentz transformation. By 1904, the Lorentz transformation had appeared in several journals and books (Darrigol 2000, 381). Einstein demonstrated the covariance of Maxwell's equations with respect to the Lorentz transformation, but the requirement of covariance of Maxwell's equations itself determines the transformations only up to a global factor (assuming linearity). Consequently, in order to derive the Lorentz transformation, imagination was required in order to set this factor equal to unity.

To this end, Lorentz (1904) advanced arguments of a physical nature, which failed to convince Henri Poincaré. If the transformation in question is to form a group, Poincaré argued, the troublesome factor can be assigned no value other than unity. Einstein took a different tack: for him, the determination of the global factor resulted from neither physical nor group-theoretical considerations, but from *kinematic* assumptions.²

He embarked upon what Martínez (2009, § 7) describes as a "tortuous" algebraic derivation of the Lorentz transformation from his kinematic assumptions, which puzzled contemporary scientists and modern historians alike. The details of Einstein's derivation have been the subject of close attention and need not be rehearsed here. Instead, I will focus on Einstein's insertion of an argument for the compatibility of his twin postulates of relativity and lightspeed invariance.³

The compatibility of Einstein's postulates of relativity and lightspeed invariance followed for Einstein from an argument which may be summarized (in slightly updated notation) as follows. Let a spherical light wave be transmitted from the

²On the assumption of linearity, see Brown (2005, 26), and for the kinematic background to Einstein's first paper on relativity, see Martínez (2009). Einstein did not let kinematics decide the matter once and for all in 1905. In a letter of September 1918 written to his friend, the anti-relativist and political assassin Friedrich Adler, Einstein considered the global factor in the Lorentz transformation to be of an empirical nature, whose value had been determined (to Einstein's satisfaction) by the results of certain electron-deflection experiments (Walter 2009, 213). Poincaré expressed his views to Lorentz by letter in May 1905; see Walter et al. (2007, §§ 38.4, 38.5).

³On the compatibility argument, see Williamson (1977). Gaps in Einstein's reasoning are apparent from a modern standpoint; see, for example, Kennedy (2005).

coordinate origin of two inertial frames designated S and S' at time $t = \tau = 0$. In system S the light wave spreads with velocity c such that the wave front is expressed as:

$$x^2 + y^2 + z^2 = c^2 t^2. \quad (1.1)$$

To obtain the equation of the wave front in frame S' moving with velocity v with respect to S , we apply a transformation of coordinates from S to S' , depending on an as-yet-undetermined factor φ , which is a function of v :

$$\xi = \varphi(v)\gamma(x - vt), \quad \eta = \varphi(v)y, \quad \zeta = \varphi(v)z, \quad \tau = \varphi(v)\gamma\left(t - \frac{vx}{c^2}\right), \quad (1.2)$$

where $\gamma = (1 - v^2/c^2)^{-\frac{1}{2}}$. Applying (1.2) to (1.1), Einstein found:

$$\xi^2 + \eta^2 + \zeta^2 = c^2 \tau^2. \quad (1.3)$$

Since (1.1) goes over to (1.3) via the transformation (1.2), Einstein observed, the light wave that is spherical in S is also spherical in S' , propagates with the same velocity c , and consequently, “our two basic principles are mutually compatible” (Einstein 1905, § 3, 901).

Einstein’s compatibility demonstration addressed one of the more immediate objections to be raised against his theory: that the propagation of light implied the existence of a substrate. This substrate, known as the ether, was common to the electron theories of Lorentz, Larmor, Bucherer-Langevin, and Abraham. Einstein’s axiomatic approach to the electrodynamics of moving bodies did not destroy the conviction that a substrate was required for light propagation. On the contrary, Einstein’s twin postulates of relativity and lightspeed invariance rendered his theory all the more suspect. Arnold Sommerfeld, for example, was impressed by Einstein’s “genial” theory, but worried that something “almost unhealthy lies in this unconstruable and intuition-free dogma.”⁴ For Tolman (1910, 28, n. 1), Einstein’s light postulate expressed “seemingly contradictory ideas” of relativity and independence of propagation velocity of light from that of its source.

Tolman’s concern over the compatibility of Einstein’s postulates stemmed in part from the fact that the propagation velocity of light is an extraordinary velocity in Einstein’s kinematics. While a spherical light wave is form-invariant for inertial frames in Einstein’s scheme, the form of other physical objects is frame dependent. A rigid sphere of matter with radius R at rest in frame S , for example, is judged by an observer in motion along the x -axis to have the flattened form of an ellipsoid of revolution with axes $(\gamma^{-1}R, R, R)$. Light waves had a special role to play in Einstein’s theory, being essentially different from other physical objects. As Einstein

⁴Sommerfeld to H.A. Lorentz, 26 Dec. 1907, in Kox (2008, § 165).

put it, the speed of light in his theory “plays the role of an infinitely great speed,” and it renders “senseless” the notion of hyperlight velocities (Einstein 1905, § 4, 903).

How did Einstein’s compatibility argument for his postulates of relativity and constant lightspeed sit with his contemporaries? At least one of Einstein’s readers, the Cambridge-trained mathematician Ebenezer Cunningham (1881–1977) was intrigued by Einstein’s approach. A student of St. John’s College, where his director of studies was the influential analytic geometer H.F. Baker, Cunningham was Senior Wrangler in 1902.⁵ Cunningham lectured on mathematics in Liverpool from 1904 and was joined there in 1906 by another Senior Wrangler (1903), Harry Bateman (1882–1946), who had studied at Trinity College. In 1907, Cunningham left Liverpool to lecture on applied mathematics at University College London, and in 1911, he returned to St. John’s as a Fellow and lecturer in mathematics.⁶

Among British theorists, relativity theory had few proponents, if any, when Cunningham first took it up.⁷ Cunningham naturally read Einstein in his own fashion, drawing on the intellectual tools at his disposal. He understood Einstein’s theory to be consistent with the existence of multiple ethers, provided that every inertial frame is associated with an ether.⁸ Inspired by Larmor’s electron theory,⁹ Cunningham’s multiple-ether view of relativity recalls the view of mechanics proposed by the Leipzig mathematician Carl Neumann. Newton’s laws of mechanics, Neumann observed, give one the freedom to consider any inertial frame to be at rest with respect to a fixed set of coordinate axes he called the “Body Alpha.”¹⁰ Neumann described the role assumed by the Body Alpha in the general theory of motion to be similar to that of the luminiferous ether in the theory of optical phenomena (Neumann 1870, 21). Views equivalent to Cunningham’s, but stripped of reference to the ether, were subsequently advanced by Minkowski (1909, 79) and Laue (1911a, 33). Cunningham’s view, based on multiple ethers, found employment throughout the 1920s, thanks to Sommerfeld’s celebrated textbook *Atombau und Spektrallinien* (Sommerfeld 1919, 319).

Cunningham’s first paper on relativity set out to overturn an objection raised by Max Abraham with respect to Lorentz’s electron theory. Abraham (1905, 205) believed that energy conservation required a fundamental modification of Lorentz’s deformable electron model, in the form of a supplemental internal,

⁵For an assessment of Baker’s rise to prominence among Cambridge geometers, see Barrow-Green and Gray (2006).

⁶See McCrea (1978), and John Heilbron’s interview with Cunningham (Heilbron 1963).

⁷A proponent of Einstein’s theory is understood here to be an individual seeking either to support or to extend any of the novel ideas contained in Einstein’s 1905 paper. On the British reception of relativity, see Goldberg (1970), Sánchez-Ron (1987), and Warwick (2003).

⁸See Goldberg (1970), and Hunt (1986).

⁹Cunningham noted a personal communication with Larmor, to the effect that while a proof of the Lorentz transformation’s validity for electron theory to second order of approximation in v/c appeared in the latter’s *Æther and Matter* (Larmor 1900), Larmor had “known for some time that [the Lorentz transformation] was exact” (Cunningham 1907, 539).

¹⁰Cunningham (1911) recalled this fact, without mentioning Neumann.

non-electromagnetic source of energy. Cunningham challenged Abraham's (frame-dependent) definition of electromagnetic momentum and found that, under the same quasistationary-motion approximation, and an alternative momentum definition, the problem vanishes. He concluded that no non-electromagnetic energy was required by Lorentz's electron model, which remained for him a possible foundation for a "purely electromagnetic theory of matter."¹¹

Along the way, Cunningham assumed that if Lorentz's deformable electron is spherical when at rest, when put in motion and measured by comoving observers, it will remain spherical. But when measured with respect to a frame at rest, the moving electron will have a "spheroidal shape as suggested by Lorentz" (Cunningham 1907, 540). Cunningham took this suggestion a step further, arguing that a light wave would appear spherical to all inertial observers, in agreement with Einstein on this point (and with reference to Einstein's relativity paper of 1905).

Next, Cunningham took an important step toward the legitimation of the concept of light-sphere covariance, as Goldberg (1970, 114) first noticed. Einstein's demonstration of the Lorentz transformation could be reduced to a handful of steps, Cunningham realized, by requiring the covariance of the light-sphere equation (1.3) with respect to this transformation. Cunningham's requirement of covariance of the light-sphere equation entailed the relativity of space and time:

For it is required, among other things, to explain how a light-wave traveling outwards in all directions with velocity C relative to an observer A , may at the same time be traveling outwards in all directions with the same velocity relative to an observer B moving relative to A with velocity v . This can clearly not be done without some transformation of the space and time variables of the two observers. (Cunningham 1907, 544)

Cunningham went on to observe that Alfred Bucherer's recent calculation of electron mass (Bucherer 1907) was mistaken, because Bucherer did not "take into account this necessary modification of coordinates."¹² Unconvinced by Cunningham's lesson, Bucherer (1908b) retorted quite rightly that light-sphere covariance was not required for the explanation of "any known fact of observation."¹³

Bucherer's complaint of inutility notwithstanding, Cunningham's clever and economical approach to deriving the Lorentz transformation represented a significant advance over Einstein's cumbersome approach via ideal rods and clocks, although he did not realize at first just what he had accomplished. This much may be gathered from his response to Bucherer's complaint, published in March 1908:

¹¹Cunningham's conclusion agrees with that reached later by Fermi; see Rohrlich (2007, 17), Janssen and Mecklenburg (2006).

¹²Cunningham (1907, 547). Both Cunningham and Planck failed to understand Bucherer's theory, which allowed for closed links between electrons; see Darrigol (2000, 371).

¹³According to Balázs (1972), Bucherer's remark shows that he was "confused about the basic problem of relativity," in that he failed to "realize the connection of this problem with the Michelson-Morley experiment and its relation to the transformation laws." Yet the Lorentz-FitzGerald contraction explains on its own the null result of the Michelson-Morley experiment, as Bucherer and contemporary theorists knew quite well.

May I explain that I did not wish to assert that [light-sphere covariance] was required by any known fact of observation, but that I took it to be involved in the statement of the principle. I may have read more into it more than was intended, but if the Maxwell equations are assumed to hold when referred, as occasion requires, to various frames of reference moving relatively to one another, the deduction cannot be escaped that the velocity of propagation of a spherical wave will be found to be exactly the same, whatever the frame of reference.

With Bucherer’s insistent prompting, Cunningham recognized not only that light-speed invariance was a convention but that he had freely interpreted Einstein’s theory.

A few months after Cunningham’s original paper appeared, Einstein employed the latter’s method in an invited review of relativity theory, making Cunningham the first British contributor to what was later known as Einstein’s theory of relativity. Einstein did not acknowledge Cunningham’s proof, however, and he may well have come up with it on his own.¹⁴

Further contributions to relativity from Cunningham and Bateman, making novel use of the light sphere, were still to come in 1908–1909. Before examining this work (in Section 1.8), it will be useful to review first the light figures produced by Poincaré and Minkowski, whose four-dimensional approach to relativity provided the basis for the later papers of Cunningham and Bateman.

1.3 Poincaré and the Lorentz Group

Poincaré (1905b) was quick to grasp the idea that the principle of relativity could be expressed mathematically by transformations that form a group. This fact had several immediate consequences for Poincaré’s understanding of relativity. Notably, following a method outlined by Lie and Scheffers (1893, 669), Poincaré identified invariants of the Lorentz transformation directly from the fact that the transformation is a rotation about the coordinate origin in four-dimensional space (with one imaginary axis). Any transformation of the Lorentz group, he noted further, may be decomposed into a dilation and a linear transformation leaving invariant the quadratic form $x^2 + y^2 + z^2 - t^2$, where light velocity is rationalized to unity. Poincaré naturally associated this quadratic form with the propagation of light and gravitational action (Poincaré 1906, §§ 4, 8).

Curiously for one who had engaged with the Riemann-Helmholtz-Lie problem of space, Poincaré avoided drawing consequences for the foundations of geometry from the “new mechanics” of the Lorentz group, with one exception. He observed that while previously, measurement of length implied the physical displacement of solids considered to be rigid,

¹⁴See Einstein (1907, § 3); reed. in Stachel et al. (1989, vol. 2, Doc. 47). Cunningham’s paper appeared in the October 1907 issue of the *Philosophical Magazine*, and Einstein’s review article was submitted for publication in Johannes Stark’s *Jahrbuch der Radioaktivität und Elektronik* on 4 December 1907.

... this is no longer true in the current theory, if we admit the Lorentzian contraction. In this theory, two equal lengths, by definition, are two lengths spanned by light in the same lapse of time. (Poincaré 1906, 132)

Light signals, in other words, were the new basis for both temporal and spatial measurement. But how was one to go about measuring lengths in a frame in motion, where measuring rods are Lorentz contracted?

Poincaré's measurement problem called for a solution, and shortly, Poincaré provided one.¹⁵ In lectures at the Sorbonne in 1906–1907, he interpreted the Lorentz transformation with respect to a geometric figure representing the wave front of an electromagnetic pulse, as judged by an observer at rest with respect to the ether. I will refer to Poincaré's figure as a "light ellipsoid," following Sommerfeld's coinage, and to a meridional section of this ellipsoid as a "light ellipse."¹⁶ The light ellipse is a staple of Poincaré's kinematics of relativity, in that he illustrated his view with this device on at least four occasions, with significant variations, during the final 6 years of his life, from 1906 to 12 July 1912. His light ellipse theory appeared three times in print during Poincaré's lifetime, in French journals of popular science, and in a book of philosophy of science.¹⁷

The origin of Poincaré's light ellipse is uncertain, but his most likely source of inspiration is a paper published by Paul Langevin in March 1905. Before discussing the latter source, however, another possible source should be mentioned: Einstein's relativity paper. As noted above, Einstein clearly expressed the spherical form of a light shell for inertial observers and used the invariance of this form under certain coordinate transformations to show the compatibility of his postulates of relativity and lightspeed invariance. Also, Einstein made an implicit distinction between an abstract geometric surface and the realization of such a surface by wave fronts as measured by inertial observers. This distinction underlies Einstein's analysis (Einstein 1905, § 8) of the energy content of a given "light complex" (*Lichtkomplex*) as measured in two inertial frames. Einstein imagined light propagating spherically in a frame S and examined the form of the surface obtained by setting to zero the time t' in the transformed wave equation. The volume enclosed by the resulting "ellipsoidal surface" (*Ellipsoidfläche*) measured in S' is greater than that enclosed by the corresponding "spherical surface" (*Kugelfläche*) measured in S , such that the energy density is less in S' than in S .

¹⁵ An alternative approach, advanced by Born (1909), involved redefining the notion of a rigid body in Minkowski spacetime. On related developments, see Maltese and Orlando (1995).

¹⁶ Sommerfeld insisted in his lectures on electrodynamics that a Lorentz transformation does not change a "Lichtkugel" into a "Lichtellipsoid" (Sommerfeld 1948, 236).

¹⁷ See the edition of Henri Vergne's notes of Poincaré's 1906–1907 lectures at the Paris Faculty of Science (Poincaré 1953) and his 1912 lectures at the *École supérieure des postes et télégraphes* (Poincaré 1913), along with the two articles (Poincaré 1908a, 1909). The article of 1908 was reedited by Poincaré in *Science et méthode* (Poincaré 1908b); the light ellipse is described on p. 239, but the diagram was suppressed from this version, presumably by the editor, Gustave Le Bon.

Einstein's focus in Section 8 of his paper was on the energy content of volumes delimited by spherical and ellipsoidal surfaces. I disagree in this instance with Arthur Miller's gloss of Einstein's argument, inasmuch as Miller identifies Einstein's equation for the ellipsoidal surface as an "ellipsoidal light pulse" (Miller 1981, 310). What Einstein wrote with respect to the equation in question was the following: "Die Kugelfläche ist – in bewegten System betrachtet – eine Ellipsoidfläche ..." (Einstein 1905, § 8). In other words, Einstein considered the energy content of an ellipsoid, and not an ellipsoidal light pulse.

More than likely, some of Einstein's contemporaries also misread Einstein's remarks on the *Ellipsoidfläche* in a moving frame and imagined an ellipsoidal light shell in a moving frame. For example, in 1912, the French polymath Maurice L  meray (1860–1926), a recognized expert on relativity theory and a former warship designer, confidently attributed a light ellipse interpretation to Einstein, only to retract his view shortly thereafter.¹⁸ While we cannot rule out the possibility that Einstein's *Kugelfläche* inspired Poincar  's light ellipse, there is a second source, which is directly linked to Poincar  's research on relativity: a paper by Paul Langevin. In the next section, I present an argument in favor of Langevin's influence on Poincar  's conception of the light ellipse.

1.4 Langevin's Electron Wake

Paul Langevin (1872–1946) was a former student of Poincar  , whose 1896 lectures at the Paris Faculty of Sciences on Sommerfeld's theory of diffraction he followed.¹⁹ Langevin had also studied at the Cavendish Laboratory, and his novel theory of the electron drew on several sources, especially Joseph Larmor's *  ther and Matter* (Larmor 1900), J. J. Thomson's *Notes on Recent Researches* (Thomson 1893), and George Searle's calculation of the energy of a Heaviside ellipsoid (Searle 1897). He introduced a distinction between the velocity fields and acceleration fields of the electron and published a graphical depiction of the velocity waves of a spherical electron in motion. This led in turn to a geometrical derivation of the field of a Heaviside ellipsoid, introduced by Heaviside in 1889, and glossed by J. J. Thomson in 1893 (op. cit.).

Langevin was concerned, as the title of his paper suggests, with the source of electron radiation and the inertia of the electron. He supposed that electron radiation was due entirely to acceleration. This stipulation allowed him to calculate the energy

¹⁸See L  meray (1912), communicated to the Paris Academy of Sciences on 9 December 1912 and the retraction (ibid., p. 1572). It is not clear whether L  meray meant to attribute a flattened light ellipsoid or an elongated light ellipsoid to Einstein. Several years later, the Swiss physicist   douard Guillaume (1921) referred to an "ellipso  de de Poincar  ." Guillaume corresponded with Einstein on this topic; see Kormos Buchwald et al. (2006, Doc. 241).

¹⁹See Langevin's notes of Poincar  's lectures, Fonds Langevin, box 123, Biblioth  que de l'  cole sup  rieure de physique et de chimie industrielle, Paris.

of an electron in uniform motion. The “electromagnetic mass” of such an electron was given to be a function of the “sillage,” or wake of the electron in motion. The wake was composed of “velocity waves” propagating, in Langevin’s picturesque language, “like the waves emanating from the front of a ship” (Langevin 1905, 171). The electron’s electromagnetic mass thus depended on the postulated charge distribution of the electron; Langevin considered both a uniform surface charge and a uniform volume charge.

The distinction between velocity and acceleration waves made here by Langevin recalls the retarded potentials introduced to electrodynamics by his former teacher Poincaré (1891), following Lorenz (1867). It recalls as well the formulation of the potentials for a moving point charge due to Liénard (1898) and Wiechert (1900). Langevin’s theory, like Lorentz’s electron theory, assumed an ether at absolute rest. According to Langevin, electrons traveled through the ether at velocities less than that of light, generating velocity waves and, in the case of non-inertial motion, acceleration waves. Both sorts of waves propagated in the ether with the speed of light, while velocity waves dissipated rapidly, such that only acceleration waves could be detected far from the electron.²⁰

1.5 Poincaré’s Light Ellipse

As a student in the mid-1890s, Langevin had followed Poincaré’s lectures on Sommerfeld’s theory of diffraction, but he did not engage personally with Poincaré until September 1904, when they were both members of the French delegation to the Congress of Arts and Sciences, held at the World’s Fair in Saint Louis. The younger man was flattered by the attention of his former teacher, as he recounted the meeting by letter to his wife back in Paris.²¹ By that time, the two men had a mutual interest in the theory of electrons, which was the topic of Langevin’s lecture in Saint Louis (Langevin 1906).

We do not know if Langevin ever discussed with Poincaré his forthcoming paper on the inertia of the electron (Langevin 1905). However, we do know that Poincaré found inspiration from the latter paper for his discovery of the Lorentz group, as it is one of the few papers cited by Poincaré (along with Lorenz 1904). Under the coordinate transformations of the Lorentz group, Poincaré demonstrated in 1905, the laws of electrodynamics retain their form. What impressed Poincaré most was not Langevin’s constant-volume model of the electron, but his explanation of the velocity and acceleration waves produced by an electron, according to which these waves propagate in free ether at the speed of light. Instead of Langevin’s model, Poincaré preferred the deformable electron model proposed by Lorenz, which had the advantage, as Poincaré proved, of preserving the principle of relativity. Poincaré (1906, 149) noticed further that by applying the Lorentz transformations

²⁰For details on Langevin’s paper, see Miller (1973).

²¹See Langevin’s notebook, box 123, and letter to his wife of 26 September 1904, box 3, Fonds Langevin, Library of the École supérieure de physique et de chimie industrielle, Paris.

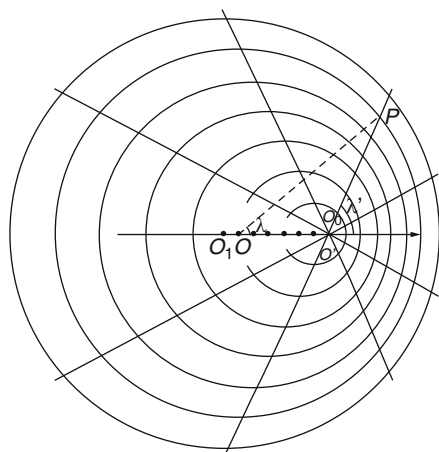
to Langevin’s acceleration waves, he could recover Hertz’s solution of Maxwell’s equations for an oscillator at rest in the absolute ether.²²

In June, 1905, Poincaré supposed that all laws of physics were likewise form-invariant with respect to the transformations of the Lorentz group, including the law of gravitation. In a letter to Lorentz announcing his discovery, Poincaré observed that the requirement of Lorentzian form invariance spelled the end of what he called the “unity of time” (Poincaré to Lorentz, in Walter et al. 2016, 2-38-3). Yet Poincaré was not ready to abandon the traditional definition of time and space in this new theoretical context. He deftly elided the question of time and space deformation in his memoir on the dynamics of the electron (Poincaré 1906) by focusing on active transformations alone (Sternberg 1986).

Questions of relativity of space and time remained on Poincaré’s mind after 1905. In his university lectures of 1906–1907, Poincaré explained how, in principle, one could measure Langevin waves and thereby determine the shape of an electromagnetic pulse generated by a source in motion with respect to the ether. According to lecture notes by a student notetaker, Henri Vergne (1879–1943), Poincaré recalled Langevin’s paper and reproduced (Figure 1.3) the latter’s illustration of the waves produced by an electron in motion (Figure 1.1).²³ He also produced a diagram of his own creation (Figure 1.2), which showed how an electromagnetic pulse was related to the Lorentz transformations. The pulse created by the point source had the form of an ellipsoid, elongated in the direction of motion of the source, with a focus collocated at the source. A section through a meridian of the ellipsoid produces the ellipse shown in Figure 1.2.

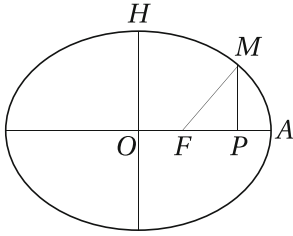
Poincaré’s diagram illustrates the Lorentz contraction, whereby all material objects contract by a Lorentz factor, but only in the direction of their motion

Fig. 1.1 Schematic diagram of an electron moving uniformly from left to right, generating velocity waves. Redrawn from Langevin (1905).



²²On Hertz’s solution, see Darrigol (2000, 251).

²³Henri Vergne, notebook 2, François Viète Center, University of Nantes.



$$\begin{aligned} \text{Lorentz factor } \gamma &= 1/\sqrt{1-v^2/c^2} \\ \text{semimajor axis } a &= OA = \gamma ct \\ \text{semiminor axis } b &= OH = ct \\ \text{eccentricity } e &= \sqrt{1-b^2/a^2} = v/c \\ \text{focal distance } OF &= \gamma vt \\ \text{apparent time } t' &= FM/c \\ \text{apparent displacement } x' &= FP \end{aligned}$$

Fig. 1.2 The light ellipse, after Vergne’s notes (2, 50). Labels H and A are added for legibility, and notation is modernized.

with respect to the ether. Commentators offer conflicting views of other aspects of Poincaré’s ellipse, and as I will show later, Poincaré himself changed his view of the ellipse around 1909.

Poincaré’s concrete model of the propagation of electromagnetic waves from a source in uniform motion merits our attention for two reasons. His light ellipse was, first of all, a graphical illustration of kinematic relations in relativity theory, the first in a long line of such techniques designed to display the relations of relativistic kinematics. Secondly, Poincaré’s theory of the light ellipse stands as the first of many attempts by physicists to reconcile an assumed Lorentz covariance of physical laws with Galilean kinematics. In particular, Poincaré’s interpretation of the Lorentz transformation contrasts sharply with the views of Cunningham and Einstein, outlined in previous sections.

The light ellipse is, at the same time, a curious historical object that has given rise to variant readings. To some extent, the lack of consensus among historians is to be expected: none of Poincaré’s four independent discussions of the light ellipse clarifies fully his protocol for measuring the dimensions of the locus of light in a moving frame. To help distinguish the various readings of Poincaré’s ellipse, let us consider three propositions:

1. The principle of relativity is valid.
2. Measurements of the light shell are performed with concrete rods by observers at relative rest with respect to rods and clocks, at an instant of *apparent time* t' indicated by light-synchronized clocks.
3. Measurements of the light shell are performed with *concrete rods in motion* by observers at *relative rest* with respect to the clocks, at an instant of *absolute time* t .

Einsteinian relativity upholds (1) and (2) only, provided that we neglect the distinction made in (2) between “apparent time” and “absolute time,” time and space being frame-dependent quantities in Einstein’s view.

The first historically motivated account of Poincaré’s light ellipse, due to Cuvaj, accepts (2), but rejects (1), in that comoving observers “will have contracted measuring sticks, in *their own frame* S' too, so that a wave-sphere (of radius ct) will appear as an ellipsoid” (Cuvaj 1970, 74, original emphasis). Thus for

Cuvaj, Poincaré’s protocol contradicts the principle of relativity, and in light of this contradiction, it is “defective.”

An alternative reconstruction of Poincaré’s measurement protocol, advanced by Wright (1975, 453) and Darrigol (1995, 41), accepts (1) and (3), such that Poincaré’s light ellipse “represents the location of a light pulse at a given value of the absolute time and for geometers belonging to a moving system” (Darrigol 1995, 41). Both Wright and Darrigol find Poincaré’s approach to be circuitous in comparison to that of Einstein. For Darrigol, Poincaré’s employment of kinematic attributes from different frames appeared “an absurdity from the Einsteinian point of view” (Darrigol 1995, 41), although he later acknowledged that Einstein, too, mixed his attributes on occasion (Darrigol 2015, note 67).

According to the reading suggested here, following Walter (2014), Poincaré originally upheld (1) and (3), but later revised his view, discarding (3) in favor of (2). Poincaré, like Einstein, considered light propagation in empty space to be the only physical phenomenon not subject to Lorentz contraction. In his first popular account of relativity theory, Poincaré (1907) drew a series of consequences for the philosophy of phenomenal space, during which he invoked a thought experiment, which proceeds as follows. Let all objects undergo the same expansion overnight; in the morning, the unsuspecting physicist will not notice any change. Poincaré likened the fantasy of an overnight spatial expansion to the relativity of moving bodies in contemporary physics, in that Lorentz’s theory admitted the contraction of bodies in their direction of motion with respect to the ether. Just as with the thought experiment, Poincaré disallowed detection of the contraction, from the assumption that instruments of measure exhibit compensating effects.²⁴

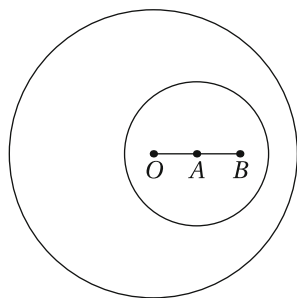
In the same vein, Poincaré admitted the principle of observational equivalence among inertial observers. He retained, however, a semantic distinction between true and apparent quantities, corresponding respectively to quantities measured in a frame at absolute rest S and those measured in frames in uniform motion with respect to the absolutely resting frame.²⁵ His definition of temporal and spatial intervals for observers in uniform motion with respect to the ether went as follows: apparent time (or equivalently, local time) is the time indicated by light-synchronized clocks at relative rest; local distance is measured by light time of flight, such that a concrete rod at rest with apparent unit length in a direction parallel to that of frame motion has true length γ .

Consequently, in an inertial frame S' , concrete measuring rods of length ℓ' contract in their direction of motion with respect to the ether frame S according to Lorentz’s formula, $\ell' = \gamma^{-1}\ell$, where $\gamma^{-1} = \sqrt{1 - v^2/c^2}$, and ℓ designates the length of the rod in a frame at rest S , v is the velocity of S' with respect to S ,

²⁴Poincaré’s fantasy was extended by Tolman (1914) via dimensional analysis, in the form of a “principle of similitude,” a view that attracted sharp criticism from Bridgman (1916).

²⁵The notion of an absolutely resting frame remained an abstraction for Poincaré. In 1912, he upheld the conventionality of spacetime and expressed a preference for Galilei spacetime over Minkowski spacetime (Walter 2009).

Fig. 1.3 A light source in uniform motion, redrawn from Vergne’s notebook (2, 50)



and c is the velocity of light, a universal constant. Observers in S' can correct for the motion-induced Lorentz contraction of their measuring rods; Poincaré put the correction factor at $5 \cdot 10^{-9}$.²⁶

In his Sorbonne lectures of 1906–1907 (mentioned above), Poincaré employed the light ellipse in pursuit of two objectives. First, he wanted to show that length and time measurements are transitive for inertial observers, transitivity being a sign of objectivity. To do so, he imagined a light source in uniform motion of velocity v that passes through the coordinate origin O at time $t_0 = 0$. At a later time $t_1 > 0$, the source reaches a point $B = vt_1$, such that the light wave originating at time t_0 and propagating in all directions with speed c has a spherical wave front of radius ct_1 . Figure 1.3, redrawn after Vergne’s notes of Poincaré’s lectures, shows a section of the surfaces of two light spheres associated with three successive positions of the source: O , A , and B . The largest light sphere has center O , and the smallest has center B , as judged by an observer at rest with respect to frame S with coordinate origin O .

According to Vergne’s notes (2, 49), Poincaré described the “measured length” of the light ellipse to be elongated in the direction of motion. I take this remark to mean that measuring rods are Lorentz contracted, such that for the resting observer, measured lengths are greater than “true” lengths by a Lorentz factor. Poincaré’s published accounts of the light ellipse do not repeat this particular description of its measured dimensions. Nonetheless, Vergne’s notes illustrate in detail Poincaré’s measurement protocol.²⁷

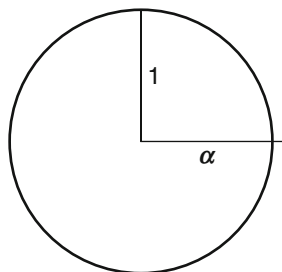
Referring to a unit circle with two segments extending from the center, as in Figure 1.4, Vergne’s notes explain the measurement procedure for an observer equipped with a ruler in motion:²⁸

²⁶See Poincaré (1901, 536), where the value is off by a factor of ten. In a later essay, Poincaré (1904, 312) supplied the “correct” value of the correction factor α for terrestrial observers and an ether at rest with respect to the Sun, where $\alpha = (\ell - \ell')/\ell = 1 - \gamma^{-1}$.

²⁷The published version of the notes differs markedly from the original, suggesting that their editor, the astronomer Marguerite Chopinet, disagreed with their content; cf. Poincaré (1953, 219).

²⁸“Alors je prends une surface rigoureusement spherique. Je la mesure avec mon mètre: dans la direction du mouvement mon mètre sera contracté de α ; sa longueur vraie sera devenue $1/\alpha$. Donc mon diamètre dans le sens du mouvement aura pour longueur mesurée α . Dans le sens

Fig. 1.4 Poincaré’s measurement scheme, redrawn from Vergne’s notebook (2, 49).



So I take a rigorously-spherical surface, and I measure it with my ruler. My ruler will be contracted by α along the direction of motion; its true length will have become $1/\alpha$. Therefore, along the bearing of motion my diameter will have the measured length α . Along the perpendicular bearing the measured length will be 1. Therefore a sphere will appear [as] an ellipsoid *elongated* along the bearing of motion. (Vergne notebook 2, 49–50, original emphasis)

Figure 1.4 shows a horizontal line segment labeled “ α ” extending from the circle center just past the circumference and a vertical segment labeled “1,” extending from the center of the circle to the circumference.

The dimensions of length measured by a comoving observer are in error due to Lorentz contraction of rulers in motion, leading Poincaré to “correct” for the contraction. Upon correction for the Lorentz contraction of rulers, Poincaré finds the “true” shape of “rigorously spherical surface” to be that of an ellipsoid of revolution, the major axis of which is aligned with the direction of motion of the observer and ruler with respect to the ether.

This measurement scheme is novel, but Poincaré went on to identify his “elongated ellipsoid” with the wave fronts of a light pulse, or what we call, for convenience, a light ellipsoid. The exact dimensions of the light ellipsoid depend on the time at which the measurement of the light locus is performed. However, the *form* of the light ellipsoid is the same for comoving observers, in that the eccentricity e is a constant that depends on frame velocity v alone, $e = v/c$ (cf. Poincaré 1908a, 393). Poincaré remarked that in a direction orthogonal to the observer’s motion, there is no motion-induced length deformation, such that the length b of the semiminor axis is $b = ct$, where t denotes “true” time, i.e., the coordinate time t_1 of the ether frame S . This remark led Poincaré to argue that apparent temporal duration is transitive for inertial frames and, ultimately, to a derivation of the Lorentz transformation.

The derivation of the light ellipse that Poincaré performed for his students proceeded as follows, based on the ellipse dimensions shown in Figure 1.2. From the diagram, Poincaré read off the standard relation for an ellipse with focus F :

$$FM + FPe = a(1 - e^2), \quad (1.4)$$

perpendiculaire la longueur mesurée sera 1. Donc une sphère paraîtra un ellipsoïde *allongé* dans le sens du mouvement.”

and then solved for t' :

$$t' = \gamma^{-1}t - vx'/c^2. \quad (1.5)$$

The latter equation shows the apparent time t' to be a linear function of apparent displacement x' , as desired.²⁹ Although Poincaré did not point this out, by simply rearranging (1.5), we obtain the transformation

$$t = \gamma(t' + vx'/c^2), \quad (1.6)$$

and upon substitution for x' , we get t' in terms of x :

$$t' = \gamma(t - vx/c^2). \quad (1.7)$$

What Poincaré *did* point out explicitly to his students (Vergne notebook 2, 51) was just this: since the difference between apparent and true time is a linear function of apparent displacement, the variable t' that appears in the Lorentz transformation is the apparent time featured in the light ellipse.

In summary, Poincaré associated during his lectures of 1906–1907 a light sphere in S of radius ct with a light ellipsoid in S' of semiminor axis length ct , and semimajor axis length γct , from the dimensions of which he derived the Lorentz transformation. Although he did not realize it, Poincaré's interpretation of the light ellipse was physically flawed, in that it ascribed to observers physical events that have no causal connection to them. The flaw can be grasped most easily by referring to a cognitive tool that was not available to Poincaré until 1908: the three-dimensional Minkowski spacetime diagram (Walter 2014).

According to the interpretation of the Lorentz group offered in Vergne's notes, the radius vector of the light ellipse corresponds to light points at an instant of ether time t . On a Minkowski diagram, the situation is described by an ellipse lying on a spacelike plane of constant time t (Figure 1.5, with the t' -axis suppressed for clarity). The ellipse center coincides with spacetime point $B = (vt, 0, t)$, and the points E, B, F , and A lie on the major axis, such that BH is a semiminor axis of length ct . The light ellipse intersects the lightcone in two points, corresponding to the endpoints of the minor axis, H and I .

In the foregoing Minkowskian representation of the light ellipse, it is plain to see that there are points on the light ellipse that lie outside the lightcone. The latter points represent locations in spacetime physically inaccessible to all inertial observers sharing a spacetime origin. In four-dimensional Minkowski spacetime,

²⁹Using the relations specified in Figure 1.2, we have

$$a(1 - e^2) = a(1 - (1 - b^2/a^2)) = a(1 - (1 - c^2t^2/a^2)) = ac^2t^2/a^2 = ct/\gamma.$$

Rearranging the latter expression in terms of t , we find $t = a\gamma(1 - e^2)/c$, and substituting the value of $a(1 - e^2)$ from (1.4), we obtain Poincaré's expression (1.5) for apparent time t' .

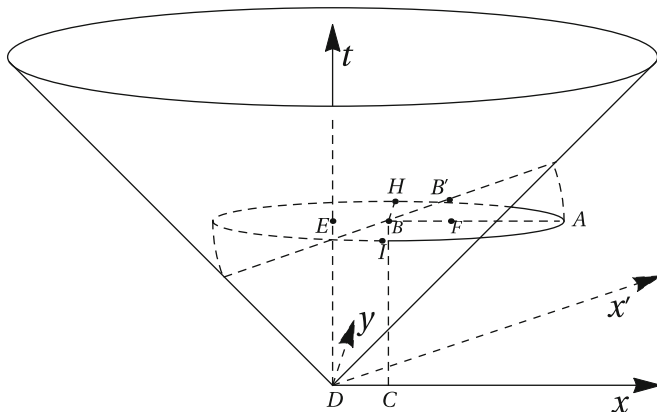


Fig. 1.5 Spacetime model of Poincaré’s light ellipse (1906) in a spatial plane ($t = \text{const.}$).

the intersection of the light sphere with center E and the light ellipsoid with center B , where E and B lie on a spacelike plane, is a circle of radius ct . On a three-dimensional Minkowski spacetime diagram, where one spatial dimension is suppressed, the corresponding circle with center E and ellipse with center B intersect in two points, labeled H and I , such that $EH = EI = BH = BI = ct$. The upshot is that Poincaré’s light ellipse model of the Lorentz group admits superluminal signals. This is certainly not what Poincaré wanted, and it may be assumed that he was not aware of the flaw in his model.

Poincaré published a popular presentation of the light ellipse in an article entitled “The dynamics of the electron” that appeared on 30 May 1908. This article recapitulates the presentation of the light ellipse found in Vergne’s notebook, and introduces a discussion of relative velocity, in which Poincaré affirms that we “must evaluate it in local time” (Poincaré 1908a, 397). While he did not explain how such an evaluation would be performed, the problem of time measurement in inertial frames was clearly posed by Poincaré. Other theorists, including Einstein and Minkowski, had posed the same question, but unlike Poincaré, they admitted that clocks in common uniform motion, synchronized by crossed light signals, are valid timekeepers.

Like Einstein and Minkowski, Poincaré came to admit that clocks in uniform motion are just as valid as clocks at rest in the absolute ether. The occasion for this step was the sixth and final lecture delivered by Poincaré in Göttingen at the invitation of the Wolfskehl Foundation, on 28 April 1909. Entitled “La mécanique nouvelle,” the lecture was the only one presented in French, as if to underline the Gallic origins of relativity theory for an audience more familiar with the theories of Einstein and Minkowski.³⁰

³⁰The context of Poincaré’s invitation to Göttingen is discussed in Walter (2018).

For his Göttingen audience, Poincaré imagined an observer in motion equipped with light-synchronized clocks and a radio transmitter-receiver.³¹ By exchanging telemetry data with a second observer in relative motion likewise equipped, the first observer comes to the conclusion that his watch is running fast. This situation corresponds to the one invoked by Poincaré in 1908, as mentioned above.³²

By allowing clocks to read local time, Poincaré was able to repair the flaw in his interpretation of the light ellipse. A few months after his lectures in Göttingen, he delivered a plenary lecture at the annual meeting of the French Association of Arts and Sciences, in Lille, on the third of August 1909. In the course of the lecture, he recalled the thought experiment from his talk in Göttingen, and noted this time that, for the two observers A and B in relative motion,³³

... a very elementary geometrical theorem shows that the *apparent* time required for light to travel from A to B , i.e., the difference between the *local* time at A when the wave leaves A and the local time at B when the wave reaches B , this apparent time, I say, is the same as if the translational motion did not exist, just as required by the principle of relativity. (Poincaré 1909, 173–174, original emphasis)

The “elementary geometrical theorem” referred to by Poincaré is just (1.4). By employing apparent time instead of ether time, Poincaré transformed in one fell swoop his light ellipse from a flawed interpretation of the Lorentz group to a model of time dilation and Lorentz contraction.

Poincaré’s employment of apparent time t' instead of ether time t , first communicated during his lecture in Lille, alters the representation of the light ellipse in a 3D-Minkowski diagram (Figure 1.6), such that the ellipse lies in a spatial plane of constant t' . The intersection of a constant-time plane $t = t_1$ with the lightcone (where $c \equiv 1$), $x^2 + y^2 - c^2t^2 = 0$ is a circle of center E and radius ct_1 in frame S , while the intersection of the lightcone with a constant-time hyperplane in S' passing through spacetime point B forms an ellipse on a spacetime diagram, corresponding to a circle of center B with respect to S' . Poincaré’s light ellipse (Figure 1.2) is identical to the intersection of the lightcone with a spacelike plane in S' passing through spacetime point B on the t' -axis. The flaw of his previous interpretation

³¹On Poincaré’s engagement with electrotechnology, and wireless telegraphy in particular, see Galison (2003), Gray (2013), and Walter (2017).

³²One may wonder why the watch in Poincaré’s thought experiment runs fast, and not slow, as would be required by time dilation in an Einsteinian or Minkowskian context. An explanation is at hand, if we focus on the first observer’s experience. At first, he believes he has a certain velocity, say 200 km/s. An exchange of telemetry data with the second observer convinces him that he is moving slower than he thought previously. One way for him to account for this revision is to admit that his watch is running fast. Other explanations for the fast watch can be imagined; see Walter (2014).

³³“... un théorème de géométrie très simple montre que le temps *apparent* que la lumière mettra à aller de A en B , c’est-à-dire la différence entre le temps *local* en A au moment du départ de A , et le temps local en B au moment de l’arrivée en B , que ce temps apparent, dis-je, est le même que si la translation n’existait pas, ce qui est bien conforme au principe de relativité.”

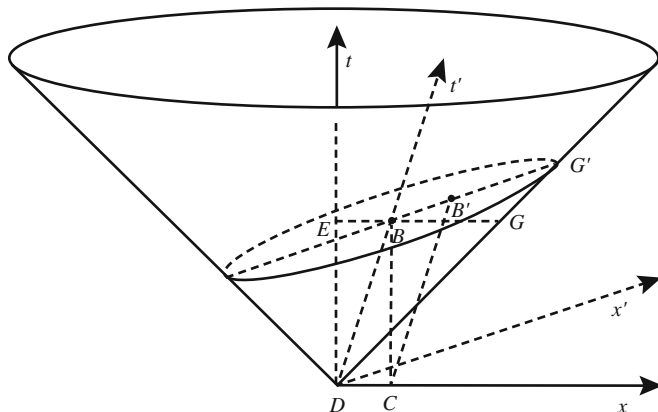


Fig. 1.6 Spacetime model of Poincaré's light ellipse (1909) in a spatial plane ($t' = \text{const.}$).

of the light ellipse (in Figure 1.5), i.e., the existence of hyperlight signals, is no longer present in the Lille interpretation, since all points of the light ellipse lie on the lightcone.

The light ellipse was not a matter of discussion for theorists during Poincaré's lifetime, and it was rarely discussed after 1912, even though Lorentz adopted the notion of a light ellipsoid in *The Theory of Electrons* (Lorentz 1909, 224). The neglect of Poincaré's light ellipse may be attributed in part to its obscure presentation in the *Revue générale des sciences pures et appliquées* (Poincaré 1908a), which was the only detailed presentation of the light ellipse to appear until 1913.³⁴ Beyond this particular case, both in France as elsewhere in Western Europe, alternatives to the Einstein-Minkowski theory were often debated, while Poincaré's theory was considered by almost no one but Poincaré.³⁵

Among electron theorists, Lorentz followed Poincaré's work more closely than others, and he applauded Poincaré's contributions, some of which he adopted, including Poincaré stress (Lorentz 1909, 213). If Lorentz was aware of Poincaré's light ellipse, he left no trace of it, while he reproduced Einstein's light-sphere derivation of the Lorentz transformations in the second edition of his *Theory of Electrons* (Lorentz 1916, 322). Lorentz did not identify the source of the derivation, which suggests that by 1916, it had lost all novelty.³⁶

Later investigators, beginning with one of Einstein's early collaborators, Guillaume (1922), invoked Poincaré's light ellipse in a quest to save the notion

³⁴An excerpt of the *Revue* article was included in Poincaré's *Science et méthode* (Poincaré 1908b), neglecting mathematical details, such as Poincaré's discussion of relative velocity.

³⁵For a sketch of the French reception of relativity, see Walter (2011).

³⁶Despite Lorentz's embrace of what Louis du Pasquier called the "principle of light-wave sphericity," the Swiss mathematician later wrote that Lorentz rejected this principle (Du Pasquier 1922, 68).

of absolute time. Guillaume's view informed the philosopher Henri Bergson's interpretation of special relativity in terms of "figures de lumière" (Bergson 1922, 133). These contributions and others are neglected here as they fall outside our temporal scope.³⁷

If Poincaré's geometric point of view is adopted, his light ellipse shows how to construe the Lorentz transformation as a rotation coupled to a dilation.³⁸ Given Poincaré's skill in conceiving intuitive models of curved space,³⁹ and in light of the fact that he interpreted the Lorentz transformation algebraically as a pure rotation in four-dimensional (3+1) space, one wonders if he considered illustrating the Lorentz transformation as a pure rotation. The latter question arises in this historical context, since Minkowski produced such an illustration just a year or so after Poincaré introduced the light ellipse. Minkowski's theory of spacetime and its relation to the light sphere and the light ellipse are taken up in the next section.

1.6 Minkowski's Lightcone

Hermann Minkowski was the first mathematician in Germany to take an interest in relativity theory. His fellow mathematicians had long abandoned research in theoretical physics, which in Germany had become the affair of specialists like Max Planck and Ludwig Boltzmann, at least since the founding of physical institutes in the 1870s (Jungnickel and McCormach 1986). But like many in mathematics, Minkowski kept abreast of research in analytical mechanics, a subject on which he lectured at Zürich Polytechnic (now the ETH), where Walter Ritz, Albert Einstein, and Marcel Grossmann were among his students. From 1902, he taught this subject and others at the Georgia Augusta University of Göttingen.⁴⁰

In Göttingen, Minkowski rejoined his friend David Hilbert and immersed himself in the activities of the local research community. The first decade of the twentieth century was a golden one for science in Göttingen, thanks in part to Felix Klein's success in attracting investments in new scientific and technical institutes from local industry and government sources and to the drawing power of the faculty. Students from Europe, Russia, the United States, and Japan came to Göttingen to hear lectures by Hilbert, Minkowski, Klein, Walther Nernst, Eduard Riecke, Woldemar Voigt, Karl Schwarzschild, Emil Wiechert, Ludwig Prandtl, and Carl Runge.⁴¹

³⁷On Guillaume's collaboration with Einstein, see Einstein's letter to Jacob Laub, 20 March 1909, in Klein et al. (1993, Doc. 143).

³⁸A displacement from one point to another on the light ellipse corresponds to a Lorentz transformation in this interpretation. The radii from a focus to any two points of the ellipse are related by a rotation and, in general, a dilation or a contraction.

³⁹On Poincaré's models of hyperbolic geometry, see Gray (1989) and Zahar (1997).

⁴⁰For background, see Walter (1999a, 2008).

⁴¹On the rise of Göttingen as a scientific center, see Manegold (1970) and Rowe (1989).

Electron theory served as a focus of many theoretical and experimental investigations undertaken by Minkowski's colleagues, although Voigt, Göttingen's chair of theoretical physics, had assumed a more critical stance. And while neither Hilbert nor Minkowski had published on questions of physics, they were keenly interested in exploring the mathematical side of electron theory, and in the summer semester of 1905, they co-directed a seminar on the subject, attended by Wiechert, the mathematician Gustav Herglotz, Born, Laue, and others (Pyenson 1979). Electron-theoretical papers by Lorentz and Poincaré figured prominently on the seminar syllabus, but their most recent publications, in which the principle of relativity and the Lorentz transformation were exploited more fully, were neglected. As for Einstein's relativity paper, it had yet to be published.

Following the electron-theory seminar, Minkowski delved into another topic of great interest to theoretical physicists: the theory of heat radiation. He lectured on recent work in this area by Planck and Nernst for the Göttingen Mathematical Society in 1906 and offered a course on the subject in the summer semester of 1907. Minkowski's course notes indicate that he was familiar with Planck's pioneering article on relativistic thermodynamics (Planck 1907), in which he praised Einstein's relativity paper. Shortly thereafter, Minkowski wrote to Einstein to request an offprint of this paper, for use in his seminar.

Einstein's achievement came to Minkowski as a "huge surprise," according to Max Born, because Minkowski did not believe Einstein possessed the mathematical background necessary to create such a theory (Seelig 1960, 45; Born 1959, 502). From the vantage point of its mathematical expression, Einstein's electrodynamics of moving bodies is all the more remarkable for its extreme simplicity. A lack of training in advanced mathematics did not constitute a serious handicap for Einstein in theoretical physics, as Hilbert and others pointed out later.⁴² In this respect, Einstein's electrodynamics of moving bodies contrasts sharply with Minkowski's memoir on the electrodynamics of moving media (Minkowski 1908), the elegant formalism of which Einstein and other physicists chose at first to ignore.

The full exploitation of light-sphere covariance required a mathematical sophistication somewhat beyond Einstein's reach in 1905. Mathematicians like Poincaré, Cunningham, and Minkowski were all in a position to explore the consequences of light-sphere covariance in their formal investigations of the principle of relativity, and all of them did so. It is far from clear, however, that Minkowski grasped the essentials of Einstein's kinematics, of which he gave a frankly distorted account in his essay "Space and Time" (Walter 1999a). As mentioned above, Einstein provided no geometrical interpretation of his kinematics or of the Lorentz transformation. In his essay, Minkowski famously illustrated his spacetime theory with geometric diagrams, and in an effort to distinguish his theory from those of Lorentz and

⁴²See Frank (1947, 206). Miller (1976, 918) emphasizes the relative simplicity of the mathematical tools deployed by Einstein in his relativity paper, in comparison to those Poincaré brought to bear on similar problems. Renn (2007, 69) observes that Einstein's uncanny aptitude for informal analysis of complex problems served him well in both special and general relativity.

Einstein, he interpreted the latter theories geometrically. Geometric reasoning carried a significant part of Minkowski's message in this work, as well as in Minkowski's earlier writings on relativity.⁴³

In one of his first attempts to provide a geometric view of the Lorentz transformation, Minkowski drew on Poincaré's observation that the Lorentz transformation corresponds to a pure rotation in four-dimensional space $(x, y, z, ct\sqrt{-1})$. During the course of a posthumously published lecture for the Göttingen Mathematical Society on 5 November 1907, Minkowski brought up the quadratic expression $x^2 + y^2 + z^2 - c^2t^2$, which he expressed in the Euclidean form $x_1^2 + x_2^2 + x_3^2 + x_4^2$, via the substitution of x_1, x_2, x_3, x_4 , for the coordinates x, y, z, ict (Minkowski 1915, 374). With this substitution, a re-expression of the laws of physics in four-dimensional terms was at hand, the premises of which Minkowski laid out in his lecture. First, however, he explored the geometry of his four-dimensional space, noting an application of hyperbolic geometry.⁴⁴ He described the hypersurface

$$t^2 - x^2 - y^2 - z^2 = 1 \quad (1.8)$$

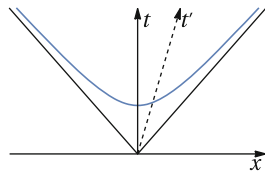
as a calibration curve of sorts, in that any line from the origin to a point on this hypersurface may be identified with the temporal axis of an inertial frame of reference. The hypersurface of equation (1.8) may also be expressed, Minkowski observed, in the form of a pseudo-hypersphere of unit imaginary radius

$$w_1^2 + w_2^2 + w_3^2 + w_4^2 = -1. \quad (1.9)$$

Both hypersurfaces (1.8) and (1.9) were known to provide a basis for models of non-Euclidean geometry (Figure 1.7).

The hypersurface (1.8) thus corresponds to the set of four-velocity vectors. Although Minkowski did not spell out the interpretation, he probably recognized that a displacement along (1.8) corresponds to a rotation ψ about the origin, such that frame velocity v is described by a hyperbolic function, $v = \tanh \psi$. However, he probably did not yet realize that his hypersurfaces represent the set of events occurring at coordinate time $t' = 1$ of inertial observers, the worldlines of which pass through the origin. According to (1.8), this time is imaginary, which may have

Fig. 1.7 A reconstruction of Minkowski's 5 November 1907 presentation of relativistic velocity space, with a pair of temporal axes, t and t' .



⁴³Minkowski's visually intuitive approach to relativity is explored at length by Galison (1979).

⁴⁴On Minkowski's use of hyperbolic geometry in this lecture, see Reynolds (1993).

obscured the latter interpretation. In fact, Minkowski did not yet possess the notion of a worldline or of proper time (Walter 2007, 217).

Sometime before the end of 1907, Minkowski discovered both worldlines and proper time, along with the lightcone structure of spacetime; he published these insights in an appendix to his theory of the electrodynamics of moving media, on 5 April 1908. The Lorentz transformation, he realized, could be written in trigonometric form, by invoking circular functions with an imaginary argument $i\psi$:

$$x'_1 = x_1, \quad x'_2 = x_2, \quad x'_3 = x_3 \cos i\psi + x_4 \sin i\psi, \quad x'_4 = -x_3 \sin i\psi + x_4 \cos i\psi,$$

where $x_4 = it$, and $c = 1$. Frame velocity v is then expressed in terms of a tangent, $v = -i \tan i\psi$. In this imaginary-angle form, the two-dimensional Lorentz transformation may be interpreted as a pure rotation about the center of a circle of imaginary unit radius. Minkowski's followers showed that in the real form, the Lorentz transformation may be construed as a displacement along a unit hyperbola. This unit hyperbola and the circle of imaginary unit radius correspond to the surfaces (1.8) and (1.9), with two spatial dimensions suppressed.

In Minkowski spacetime, Einstein's light-sphere covariance gives way to lightcone covariance. Minkowski interpreted Einstein's expression (1.1) for a light sphere as the equation of a lightcone in spacetime. Whereas both Einstein and Poincaré understood light waves in empty space to be the only physical objects immune to Lorentz contraction, Minkowski saw that when light rays are considered as worldlines, they divide spacetime into three regions, corresponding to the spacetime region inside a future-directed ($t > 0$) hypercone ("*Nachkegel*"), the region inside a past-directed ($t < 0$) hypercone ("*Vorkegel*"), and the region outside any such hypercone pair. The propagation in space and time of a spherical light wave is described by a hypercone, or what Minkowski called a lightcone ("*Lichtkegel*").

One immediate consequence for Minkowski of the lightcone structure of spacetime concerned the relativity of simultaneity. In a section of his paper on the electrodynamics of moving media entitled "The concept of time," Minkowski (1908, § 6) showed that Einstein's relativity of simultaneity is not absolute. While the relativity of simultaneity is indeed valid for two or three simultaneous "events" (*Ereignisse*), the simultaneity of four events is absolute, so long as the four spacetime points do not lie on the same spatial plane.⁴⁵ Minkowski's demonstration relied on the Einstein simultaneity convention, and employed both light signals and spacetime geometry, but not the light sphere. His result showed the advantage of employing his spacetime geometry in physics, and later writers, including Poincaré, appear to have agreed with him, by considering the discovery of the existence of a

⁴⁵"Werden jedoch vier Raumpunkte, die nicht in einer Ebene liegen, zu einer und derselben Zeit t_0 aufgefaßt, so ist es nicht mehr möglich, durch eine Lorentz-Transformation eine Abänderung des Zeitparameters vorzunehmen, ohne daß der Charakter der Gleichzeitigkeit dieser vier Raum-Zeitpunkt verloren" (Minkowski 1908, 69).

class of events for a given observer that can be the cause of no other events for the same observer as a consequence of spacetime geometry (Walter 2009, 210).

Physicists in Germany quickly seized upon Minkowski's electrodynamics of moving media, but as mentioned above, they stripped it of the four-dimensional formalism in which it had been dressed by its inventor. In what became the standard response to Minkowski's electrodynamics of moving media, both in Germany and abroad, Minkowski's former students Einstein and Jacob Laub recast Minkowski's four-dimensional expressions in terms of ordinary vectors. In 1908, outside of Göttingen and Cambridge, theorists saw no use at all for a four-dimensional approach to physics.

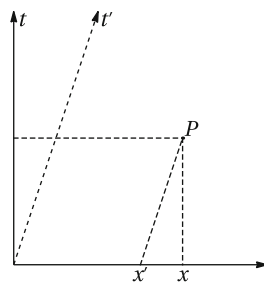
One imagines that for Minkowski, this was a vexatious state of affairs. According to his former student Max Born, Minkowski always aspired

to find the form for the presentation of his thoughts that corresponded best to the subject matter. (Born 1914)

The form Minkowski gave to his theory of moving media had just been judged unwieldy by his readers, and in the circumstances, decisive action was called for if his formalism was to survive at all. In September 1908, he took such action, by affirming the reality of the four-dimensional "world" and its necessity for physics (Walter 2010). His celebrated lecture "Raum und Zeit," delivered at the annual meeting of the German Association of Scientists and Physicians in Cologne, offered two diagrammatic readings of the Lorentz transformation, one attributed to Lorentz and Einstein, the other to himself.

The first of these two readings was supposed to represent the kinematics of the theory of relativity of Lorentz and Einstein. In fact, Minkowski's reading captured Lorentzian kinematics, but distorted Einsteinian kinematics, prompting corrective action from Philipp Frank, Guido Castelnuovo, and Max Born.⁴⁶ The idea stressed by Minkowski was that in the (Galilean) kinematics employed in Lorentz's electron theory, time being absolute, the temporal axis on a space-time diagram may be rotated freely about the coordinate origin in the upper half-plane ($t > 0$), as shown in Figure 1.8. The spatial position of a point P may be described with respect to frames

Fig. 1.8 A reconstruction of Minkowski's depiction of the kinematics of Lorentz and Einstein, after Born (1920).



⁴⁶See Born (1909, 9; 1959, 503). For further references and details on Minkowski's distortion and its reception, see Walter (1999a).

S and S' , corresponding to the coordinate axes (x, t) and (x', t') , respectively, according to the coordinate transformation: $x' = x - vt$, $t' = t$.

In contradistinction to the latter view, the theory proposed by Minkowski required a certain symmetry between the spatial and temporal axes. This constraint on symmetry itself was sufficient for a geometric derivation of the Lorentz transformation. Although Minkowski described his spacetime diagram as an illustration of the Lorentz transformation, he did not spell out the interpretation in detail. Nonetheless, Minkowski did provide a geometric derivation of the Lorentz transformation at some point, as attested by an autograph slide in Minkowski's Nachlass, which may have been projected during the lecture he delivered to the German Association of Scientists and Physicians in Cologne.⁴⁷

While Minkowski acknowledged Einstein's critique of absolute time, he considered that the concept of a rigid body – upon which Einstein had based his relativistic kinematics – made no sense in relativity theory (Minkowski 1909, 80). Similarly, Poincaré deemed that measurement in relativity theory could no longer rely on the displacement of rigid bodies, which were replaced for the purpose of measurement in Lorentz's theory by light time of flight (Section 1.3). For Poincaré and Minkowski, Einstein's foundation of relativistic kinematics on the behavior of ideal clocks and rigid rods did not sit well at all with the Lorentz deformation of displaced solids. They did not appeal to the kinematics of rigid bodies to derive the Lorentz transformation, but affirmed the principle of relativity, and required that the transformations of coordinates between inertial frames form a group.⁴⁸ For Minkowski, Lorentz contraction of electrons was a direct consequence of the geometry pertaining to this group (Figure 1.9).

The latter consequences were displayed by Minkowski on a spacetime diagram and elaborated upon by Sommerfeld on the occasion of a reedition of Minkowski's Cologne lecture (Sommerfeld 1913). Minkowski's spacetime diagram thus offered a novel means of understanding the strange consequences of Einstein's kinematic assumptions. However, the spacetime diagram was understood by some physicists to lend support to an ether-based outlook, as shown by Emil Wiechert's contributions (discussed in Section 1.9).

⁴⁷NSUB Handschriftenabteilung. The demonstration missing from the published text of Minkowski's lecture was later supplied by Arnold Sommerfeld, in an editorial note to his friend's lecture. The annotated version of the lecture appeared in an anthology of papers on the theory of relativity edited by Blumenthal (1913). According to Rowe (2009, 37), Sommerfeld was the driving force behind the latter anthology.

⁴⁸As seen above, Poincaré also derived the Lorentz transformation from the assumption of Lorentz contraction of concrete rods, and the isotropy of light propagation for inertial observers. He later considered (apparent) time deformation as a consequence of the principle of relativity and Lorentz contraction; see (Poincaré 1913, 44).

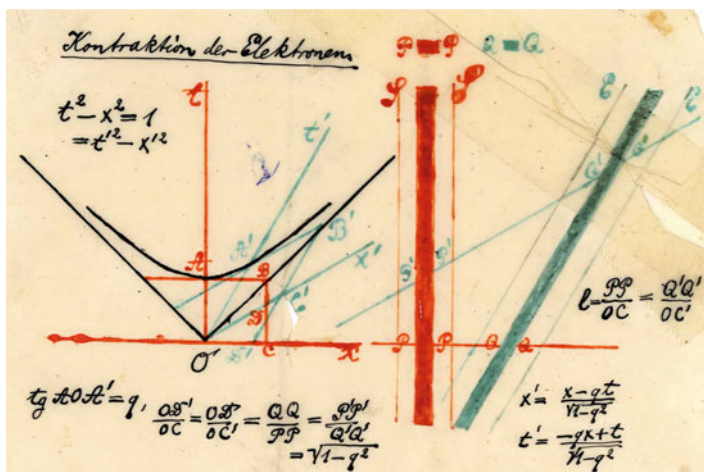


Fig. 1.9 An autograph, hand-colored transparency of Minkowski's geometric derivation of the Lorentz transformation, probably from the Cologne lecture of 21 September 1908. Courtesy of the Niedersächsischen Staats- und Universitätsbibliothek, Handschriftenabteilung.

1.7 Alfred A. Robb: Repurposing the Lightcone

A physicist trained in Belfast, Cambridge, and Göttingen, Alfred A. Robb (1873–1936), found the means in Minkowski's spacetime geometry to realize an “optical geometry of motion,” in which he could dispense with Einstein's ideal clocks and rigid rods (Robb 1911). Robb, described by Larmor (1938, 320) as an “unremarkable” graduate of St. John's College, Cambridge, was ranked fifty-second (*ex aequo*) in the 1897 Mathematical Tripos.⁴⁹ He went on to write a theoretical study of the Zeeman effect in Göttingen under W. Voigt's direction, published in the *Annalen der Physik* (Robb 1904), after which he returned to St. John's, and joined the Cambridge Philosophical Society.⁵⁰

In his doctoral thesis, Robb took up one of the more puzzling problems facing physicists in the early twentieth century: to explain the patterns of magnetic splitting of atomic spectral lines, known then as the complex Zeeman effect. Starting from Lorentz's Nobel Prize-winning theory of doublet and triplet lines (Lorentz 1897), Robb introduced elastic forces between electron pairs, triplets, and quadruplets. To obtain agreement with observation, he introduced a geometric constraint, requiring electrons to oscillate on the surface of a cone. As he wrote

⁴⁹Tanner (1917, 571). I thank J. Barrow-Green for pointing me to this source.

⁵⁰Robb was admitted to the Society on 27 Nov. 1905 (*Proceedings of the Cambridge Philosophical Society* 16, 1912, p. 16).

to Larmor, the “restrictions are so peculiar that one may be inclined to doubt the theory,” and indeed, Robb’s scheme was later described by Lorentz as both “very ingenious” and “so artificial.”⁵¹

Much like his theory of the complex Zeeman effect, Robb’s optical geometry was both ingenious and unattractive to physicists. Yet Robb’s geometry laid the groundwork for a theory of time and space that was later hailed by the likes of Weyl (1922, 209). To build his optical geometry, Robb borrowed some basic insights from Minkowski and transformed them as needed. For example, he employed Minkowski’s trigonometric definition of velocity v , in a real hyperbolic form, such that $v = \tanh \omega$, and called ω the “rapidity” of the particle.⁵² Taking a cue from the Minkowski spacetime diagram, Robb described particle velocity with respect to the index axis z via the relation $\tan \gamma = \tanh \omega$ and expressed the lightcone in terms of orthogonal axes xyz :

$$x^2 + y^2 - z^2 = 0, \quad (1.10)$$

where the z -axis represents a temporal index, the vertex coincides with the origin, and the speed of light is unity.

For purposes of illustration, Robb followed Minkowski’s convention on units, such that the path of light in vacuum is described for any inertial observer by a line forming an angle of 45° with the z -axis. In the place of Einstein’s notion of distant simultaneity, Robb introduced a more restrictive definition, whereby the emission or reception of two or more light signals is simultaneous if and only if it is observed at a single spatial location at a single instant of time by a colocated, inertial observer. Simultaneity is an absolute notion in Robb’s scheme, and the distance to a particle of matter in arbitrary motion is determined by round-trip light time of flight between the inertial observer and the particle.

Light rays play a fundamental role in Robb’s geometry of phenomenal space, as the title of his booklet suggests. Issues of clock synchronization do not arise here, nor is there any question of transforming measured quantities. However, Robb was careful to show that according to his theory, lengths of material bodies “appear to be shortened in the direction of motion,” a result in agreement with other relativity theories.⁵³ Moreover, to demonstrate this result, Robb implicitly borrowed

⁵¹Robb to Larmor, 6 March 1904, Larmor Papers, St. John’s College Library; Lorentz (1909, 115). Voigt sent Lorentz a copy of Robb’s dissertation; see Lorentz to Voigt, 18 Dec. 1904, in Kox (2008, § 121).

⁵²In a letter to Larmor of 18 Jan. 1911, the Cambridge mathematician A. E. H. Love wrote that he had “noted explicitly in writing” to Robb that one of his formulas was from Lobachevski geometry, and that “space might be saved by bringing this fact in” (Larmor Papers, St. John’s College Library). On Robb’s use of hyperbolic geometry, see Walter (1999b).

⁵³For Robb the “appearance” of contraction was a necessary consequence of light time-of-flight measurements. Robb, Einstein, and their contemporaries focused on the instantaneous form of moving objects, in an approach distinct from the one adopted in the late 1950s. The latter studies

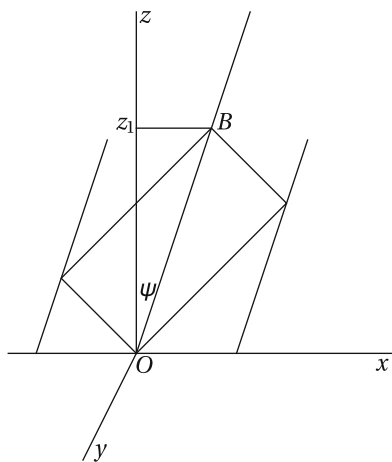
Poincaré’s idea of using a light shell as a metric surface. Yet Robb openly distanced himself from Poincaré’s conventionalist philosophy of geometry:

Speaking of the different “Geometries” which have been devised, Poincaré has gone so far as to say that : “one Geometry cannot be more true than another; it can only be more convenient.” [...] In reply to this; it must be remembered that the language of Geometry has a certain fairly well defined physical signification which *in its essential features* must be preserved if we are to avoid confusion.⁵⁴

From the latter remark, Robb’s philosophical position appears closer to Einstein’s contemporary view of a physical geometry realized by ideal rigid rods and clocks than to Poincaré’s conventionalist doctrine, which ruled out any empirical determination of the geometry of phenomenal space. But as mentioned above, Robb did not admit Einstein’s distant simultaneity, rigid rods, or ideal clocks.

Robb’s philosophy of geometry was an innovative response to the relativity theories of Einstein, Poincaré, and Minkowski, which he developed from around 1910 until the end of his life.⁵⁵ In his first publication on geometry (Robb 1911), Robb’s philosophy found expression in an original analysis of the form of a reflected light shell for an observer in motion. His approach to this problem employed a diagram of a three-dimensional space, redrawn here as Figure 1.10; it may be summarized briefly as follows. An observer in uniform motion along the x -axis with velocity $v = \tan \psi$ transmits a number of light signals in the xy -plane at an instant of time with index $z_0 = 0$. These signals are reflected from a ring of comoving particles surrounding the observer in such a way that the signals arrive at the observer’s location at point B at a single instant of time of index $z_1 > z_0$.

Fig. 1.10 A ring of particles in uniform motion in xyz -space, after Robb (1911). Labels z_1 and B are introduced for clarity.



characterized what Penrose (1959) referred to as the “photographic” appearance of a moving object.

⁵⁴Robb (1911, 1), original emphasis. Cf. Poincaré, *Science and Hypothesis* (Poincaré 1905a, 50).

⁵⁵For appreciations of Robb’s geometry, see Briginshaw (1979) and Cat (2016).

Neglecting one spatial dimension, Robb's diagram shows a future lightcone from the origin in xyz -space that intersects with a past lightcone with vertex at point $B = (z_1 \tan \psi, 0, z_1)$. The intersection of the two lightcones defines an ellipse on an oblique plane, which is not illustrated in Robb's diagram, but which contains a diagonal of the rectangle in Figure 1.10, and forms an angle ψ with the x -axis. By way of comparison, in Minkowski three-dimensional spacetime, the latter plane corresponds to a spacelike plane of an observer in motion with velocity v .

Iteration of the signaling process produces a representation of an elliptic cylinder of axis OB , the equation for which Robb derived. From the perspective of an observer at rest with respect to the origin, Robb argued, the "apparent form of this ring of particles" is given by a section of the cylinder in the xy -plane. In other words, the apparent form, for an observer in the rest frame, of the (reflected) light shell of an observer in motion is an ellipse of eccentricity $\sqrt{1 - v^2}$, the minor axis of which is aligned with the ring's direction of motion. Naturally, Robb concluded that the length of objects in motion, when measured by light time of flight, appears to a resting observer to be contracted in the direction of motion.

Although Robb did not say as much, his observer in motion is in a position to conclude that the reflecting ring of particles forms a circle. If Robb's observer assumes, with Poincaré, that her concrete rods are Lorentz contracted, she may correctly infer that her light shell is an elongated ellipsoid, the dimensions of which agree, moreover, with Poincaré's light ellipsoid. Robb's measurement protocol, however, featured no such concrete rods. Furthermore, unlike Poincaré, Robb admitted no privileged frame of reference. Consequently, Robb could no more uphold Poincaré's homotheticity of light ellipsoids than he could affirm Einstein's covariance of light spheres. His preferred figure of light was the lightcone.

In late 1910, Robb had submitted a like-titled work for publication in the *Proceedings* of the London Mathematical Society (LMS).⁵⁶ One of the Society's two secretaries (along with J. H. Grace), A. E. H. Love asked Robb to "withdraw his paper temporarily," in order to address the criticisms of a referee. Robb appears to have complained about Love's request to his former teacher at St. John's, and LMS council member, Joseph Larmor. In response to Larmor's query, Love wrote that a referee had found the geometrical part of Robb's paper to be "extremely illogical" and had recommended rejection. Love noted that the theory of relativity also entered into the referee's assessment: Robb's time index formula was "suggested by Einstein's work," and furthermore, the referee felt "it might be necessary to adopt Einstein's assumptions in order to have some basis for Robb's formula."⁵⁷ The LMS reviewer's critical assessment of Robb's work was echoed in softer terms by the Cambridge logician P. E. B. Jourdain (1879–1919). According to Jourdain's published abstract, the formulas in Robb's booklet agreed with those of Einstein,

⁵⁶LMS Council Minutes, 10 Nov. 1910, LMS archives.

⁵⁷Love to Larmor, 18 Jan. 1911, op. cit. Sedleian Chair of Natural Philosophy at Oxford since 1899, Love was Secretary (i.e., managing editor) of the LMS from 1890 to 1910.

Minkowski, and Sommerfeld, but the concepts he employed, including that of the index of a particle, were such that no summary could be provided.⁵⁸

From these assessments, it appears that Robb's theory was seen in England as a confusing, mathematically inept variant of Einstein's theory. The decision by the LMS council to follow Grace and Love's recommendation against publication of Robb's manuscript meant his theory would not benefit from a stamp of authority from Britain's leading mathematicians and theoretical physicists.⁵⁹ Robb's rejected manuscript then became a booklet, the preface to which, dated 13 May 1911, suggests that its author was still smarting from the LMS council's negative decision:

From the standpoint of the pure mathematician *Geometry* is a branch of *formal logic*, but there are more aspects of things than one, and the geometrician has but to look at the name of his science to be reminded that it had its origin in a definite *physical* problem. That problem in an extended form still retains its interest.

The italics in the passage above are Robb's, underlining the triad: geometry–logic–physics. Beyond the expected retort to the censorious pure mathematician, Robb's preface affirmed his identity as a “geometrician” and his consequent right to “extend” the domain of application of geometry beyond that of both logic and the measurement of length intervals in the phenomenal space of physics.

From the LMS council's rejection of Robb's theory of relativity, one gathers that this theory had its detractors. But as mentioned above, the theory had its admirers, as well. One of these admirers was Ludwig Silberstein (1872–1948). A former doctoral student of Max Planck in Berlin, Silberstein wrote *The Theory of Relativity* (Silberstein 1914), one of the first two textbooks on the subject to be published in England, with Cunningham's *The Principle of Relativity* (Cunningham 1914). A lecturer in mathematical physics in Rome since 1903, Silberstein based his textbook on lectures delivered at University College London in 1912–1913. In his preface, Silberstein explained his wish “to trace the connexion of the modern theory with the theories and ideas that preceded it.” The modern theory Silberstein referred to here was essentially that of Einstein and Minkowski.

In a chapter of his textbook entitled “Various Representations of the Lorentz Transformation,” Silberstein recommended Minkowski diagrams, described as “very advantageous, especially for the trained geometer of our days” (Silberstein 1914, 131). His overview of the “geometric representation” of the Lorentz transformation began with a two-dimensional spacetime diagram, illustrated by a figure showing two pairs of coordinate axes, the conjugate hyperbolas $x^2 - c^2t^2 = -1$ and $x^2 - c^2t^2 = 1$, and their asymptotes. He recalled that for any real number κ , the two families of hyperbolas $x^2 - c^2t^2 = -\kappa$ and $x^2 - c^2t^2 = \kappa$ are Lorentz covariant. Extending his arguments to three spacetime dimensions, and then four, Silberstein observed (p. 139) that the spacelike hypersurface ($t = 0$) intersects the hyperboloid $x^2 + y^2 + z^2 - c^2t^2 = 1$ in a unit sphere, $x^2 + y^2 + z^2 = 1$. A non-zero

⁵⁸ *Jahrbuch über die Fortschritte der Mathematik* 43, 1911, p. 559. A succinct summary of Robb's index concept is provided by Barrow-Green and Gray (2006).

⁵⁹ LMS Council Minutes, 9 Feb. 1911, LMS archives.

rotation of this hypersurface about the origin in a plane orthogonal to the t -axis cuts the hyperboloid in an ellipsoid, resulting in a primed space, $x'y'z'$, and an assorted orthogonal axis, t' . Silberstein continued:

Take the semi-diameters of this ellipsoid as the new units of length measured from the origin along any direction in the $x'y'z'$ -space. Then the Lorentz transformation, from S to S' , will be completed, and the new metric surface which, from the S -point of view, is an ellipsoid of revolution will for the S' -standpoint become a sphere, $x'^2 + y'^2 + z'^2 = 1$.

According to Silberstein's analysis, the intersection of a t' -constant hypersurface with a Lorentz-covariant hyperboloid in spacetime is an ellipsoid of revolution in the S -frame and a sphere in the S' -frame.

1.8 Applications of the Light Sphere

Minkowski's spacetime theory was understood to be consistent with Einstein's concept of light-sphere covariance, the latter being considered both as a special case of Lorentz covariance of the laws of physics and as a mathematical theorem. The *figure* of a light sphere, however, was never discussed by Minkowski. Nonetheless, physicists like Wiechert (1911, 691) understood the derivation of the Lorentz transformation from the form invariance of the light-sphere equation to be the true "point of departure" of Minkowski's spacetime theory. Such a reading suggests that Einstein's light sphere prepared scientists for the formal requirement of Lorentz covariance for the laws of physics, as manifested in Minkowski's theory and as realized in four-dimensional vector and tensor algebras by Sommerfeld, Abraham, Gilbert Newton Lewis, Laue, and others.

Minkowski employed the equation of a light sphere in his representation of the Lorentz transformation by postulating the invariance of the quadratic form:

$$-x^2 - y^2 - z^2 + t^2, \quad (1.11)$$

where the velocity of light is rationalized to unity (Minkowski 1908, 66). Next, invoking the substitution x_1, x_2, x_3, x_4 for coordinates x, y, z, it , Minkowski expressed the general Lorentz transformation in terms of a 4×4 coefficient matrix A ,

$$A = \begin{vmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{vmatrix}, \quad (1.12)$$

with determinant unity such that

$$x_h = \alpha_{h1}x'_1 + \alpha_{h2}x'_2 + \alpha_{h3}x'_3 + \alpha_{h4}x'_4 \quad (h = 1, 2, 3, 4). \quad (1.13)$$

Cunningham was struck by Minkowski's equation (1.12) and by the fact that Minkowski's restriction on the determinant could be relaxed while preserving the form of the wave equation.⁶⁰ The latter insight was exploited in the papers he and Bateman published on the conformal transformations of Minkowski spacetime in 1909–1910. In addition to the form (1.12), Cunningham (1910, 79) acknowledged Minkowski's interpretation of the Lorentz transformation in relation to the light-wave equation:

It has been pointed out by Minkowski that in a space of four dimensions in which the coordinates are $(x, y, z, ct\sqrt{-1})$, the geometrical transformations employed by Einstein, is simply a finite rotational displacement of the whole space about $y = 0, z = 0$. The equation $\nabla^2 V = 0$ [...] is known to be invariant for such a transformation. But this equation is invariant for a larger group of transformations than that of rotations, viz., for the group of conformal transformations in the four dimensional space, which, as is known, is built up out of inversions with respect to the hyperspheres of the space. (Cunningham 1910, 79)

Cunningham noted further (p. 80) that the hyperspace (x, y, z, ict) is conformal to the hyperspace (X, Y, Z, icT) in virtue of the form invariance of the light-sphere equation. He was, however, not the first to notice the conformal covariance of the wave equation in Minkowski spacetime. For the latter insight, Cunningham acknowledged a remark made to him by his former colleague in Liverpool, Harry Bateman.⁶¹

Following his success in the Mathematical Tripos, Bateman undertook two years of postgraduate study in Paris and Göttingen, then major centers for experimental and theoretical research on electrons. A central topic of discussion in mathematical physics at the time, the electron theories of Lorentz and Larmor were introduced to French readers by Poincaré and Liénard starting in 1897 (Buchwald 1985). Similarly, in Göttingen, Emil Wiechert, Karl Schwarzschild, and Max Abraham contributed to electron theory, while the mathematicians Hilbert and Minkowski co-ed seminars on electron theory and electrodynamics in 1905 and 1907, respectively (Pyenson 1979). When Bateman studied in Göttingen, he was particularly impressed by Hilbert's approach to integral equations, a subject he taught at Cambridge in 1908.⁶²

Returning to England in 1906, Bateman joined Cunningham as a lecturer at the University of Liverpool. There he applied W. Thomson's method of inversion to geometrical optics and found the form of the differential equation for light-wave propagation to be preserved under conformal transformations of four-dimensional (Minkowski) space, much as Minkowski had observed with respect to the transformations of the (inhomogeneous) Lorentz group.⁶³ Bateman also remarked that his

⁶⁰Cunningham (1914, 87–89); for an analysis of the procedure, see Newman and Price (2010).

⁶¹See Cunningham (1910, 79). As for Bateman, he credited Cunningham with the discovery of the conformal transformations of the equations of electrodynamics; see Bateman (1910c, 224).

⁶²*L'Enseignement mathématique* 10 (1908), 336; Bateman to Hilbert, 1909, Nachlass Hilbert 13, Handschriftenabteilung, NSUB Göttingen.

⁶³See Minkowski (1909), where the Lorentz transformation is attributed to a paper published in 1887 by Voigt. Minkowski described the covariance of the differential equation of light-wave

method gave rise to a “geometrical construction” in ray optics, whereby a sphere of radius ct and center (X, Y, Z) corresponds to an inverse sphere of radius cT with center (x, y, z) . In other words, light spheres transform into inverse light spheres.⁶⁴

Cunningham and Bateman were atypical in their rapid assimilation of Minkowski’s four-dimensional approach to electrodynamics, a fact which may be attributed in part to local factors, including their training in mathematics. In particular, the technique of conformal transformation was part of the Wrangler’s mathematical arsenal from at least the mid-1890s, and studies of the conformal group in space of n dimensions ($n > 2$) were familiar in Cambridge at the turn of the century.⁶⁵ In Bateman’s case, postgraduate studies in Göttingen in 1906 afforded him personal acquaintance with Hilbert and Minkowski, both of whom were instrumental in the elaboration and diffusion of four-dimensional approaches to physics. In a fashion similar to that of the latter pair, but on a smaller scale, Cunningham and Bateman employed and promoted four-dimensional techniques: Bateman (1909) published Maxwell’s equations and Lorentz’s microscopic equations in four-vector form, while Cunningham’s treatise on relativity (Cunningham 1914) featured an introduction to four-dimensional vector calculus.

Some of the earliest contributions to relativity theory are due to one of the youngest relativists: Max Laue (1879–1960). A former doctoral student, then assistant to Max Planck, upon whose suggestion he wrote a doctoral dissertation (Laue 1903) on interference phenomena, Laue first heard of Einstein’s relativity paper in Planck’s colloquium. As he recalled later, he found that Einstein’s paper presented “epistemological difficulties” that he was initially unable to master.⁶⁶

Over the next 5 years, Laue came to master a few of the difficulties presented by Einstein’s theory, beginning with the derivation of the Fresnel drag coefficient from the velocity addition formula (Laue 1907). He adopted a comparative approach to the electrodynamics of moving bodies, publishing a series of papers evaluating the various theoretical options. One of his first contributions compared the electron theories of Abraham and Lorentz to what he called Einstein’s “theory of relativity” (Laue 1908, 838). The differences between the two electron theories, Laue found,

propagation as the “impetus and true motivation” for assuming the covariance of all laws of physics with respect to the transformations of the Lorentz group (p. 80).

⁶⁴See Bateman (1908, 629), read 8 Sept. 1908. No mention is made in this paper of the source of the transformations, but a subsequent work by Bateman credits Cunningham with the “discovery of the formulæ of transformation in the case of an inversion in the four-dimensional space” and cites papers by Hargreaves and Minkowski employing a four-dimensional space with one imaginary axis (Bateman 1909, 224, communicated 9 Oct. 1908). Minkowski’s paper (Minkowski 1908, published 5 April 1908) was cited by both Cunningham and Bateman. Remarkably first by Whittaker (1951, vol. 2, 195), the significance of Minkowski’s spacetime theory for the contributions of Cunningham and Bateman is contested by Warwick (2003, 423 n. 49). On the “light-geometric approach” to the foundations of relativity by Cunningham and Bateman, see Jammer (1979, 222).

⁶⁵For example, see Warwick (2003, 421) and Bromwich (1901).

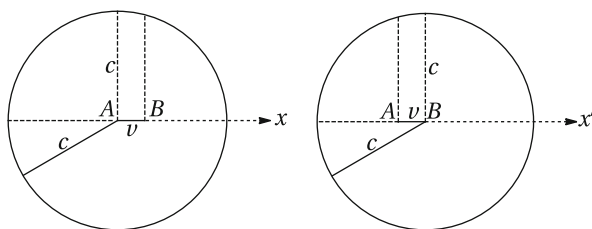
⁶⁶von Laue (1961, XVIII–XXI); von Laue to Margot Einstein, 23 Oct. 1959, cited by Holton (1965, 39).

were too small to matter as far as the radiation from a charged particle in motion was concerned, but there was an advantage in adopting Einstein's theory, in that it was "much simpler" to solve the latter problem (*ibid.*). When in 1911 Laue extended Minkowski's four-dimensional approach to the dynamics of matter via the formal concept of a "world tensor," he was able to shed new light on the nature of Poincaré's hypothetical binding potential, later known as "Poincaré pressure," and on the null result of the Trouton-Noble experiment of 1903.⁶⁷

Recognizing Laue's skill in addressing the questions posed by relativity theory, the publishing house of Vieweg asked him in 1910 to write what was to become the first textbook on relativity.⁶⁸ There Laue identified Einstein as the principal founder of the theory of relativity. He did so, however, while expounding a four-dimensional vector calculus he attributed to Minkowski, but which owed more to Sommerfeld's formalism.⁶⁹ Laue's text thus helped established Einstein as a leading theorist in the new field of relativity and to promulgate four-dimensional tensor calculus.⁷⁰

One of the results Laue included in his textbook was the light-sphere-based illustration of the relativity of simultaneity. Laue's argument and illustration drew on an idea expressed earlier by Planck in lectures delivered at Columbia University in 1909 and published the following year. Planck wanted to convey graphically what he called the "new difficulty" introduced by the principle of relativity, concerning the propagation velocity of light in the ether (Planck 1910, 113). To do so, he referred to two diagrams (see Figure 1.11), representing a section of a light sphere for observers *A* and *B*, respectively, with relative velocity *v*. Taken separately, each of the two diagrams suggests that light isotropy is valid only for observers at rest, since apparently, only such observers will find themselves at the center of the light sphere. Planck stressed, however, that no known physical phenomena distinguished

Fig. 1.11 Meridional section of a light sphere for an observer *A* at rest (left) and an observer *B* at rest (right), after Planck (1910, 114, 119).



⁶⁷Laue (1911b); Janssen and Mecklenburg (2006).

⁶⁸von Laue (1952).

⁶⁹See Max Born's review in *Physikalische Zeitschrift* (Born 1912).

⁷⁰On Laue's portrayal of Einstein's contribution, see Staley (1998). Laue's contributions to relativity are detailed by Norton (1992) and Rowe (2008).

the two frames and that the difficulty could be overcome by admitting, with Einstein, the Lorentz covariance of the laws of physics (Planck 1910, 121).⁷¹

Laue took a different approach, by adapting Planck's light figures, in order to address the notion of relative simultaneity. Repeated in six editions by 1956, Laue's light figure became a staple of presentations of relativity theory. Pared to essentials, his argument (Laue 1911a, 34) focused on the simultaneity relation as judged by two observers, respectively, at rest and in uniform motion. A "short light signal" is emitted in all directions by a source at rest at a "material point" A in inertial frame S at time $t_0 = 0$. If the origin of coordinates is fixed at point A , then at time $t > 0$, the light signal reaches the points described by the equation:

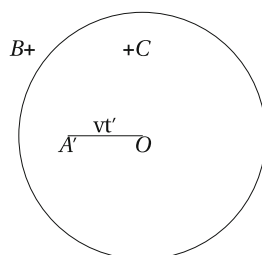
$$x^2 + y^2 + z^2 - c^2t^2 = 0. \quad (1.14)$$

Let two "material points" B and C at rest in S be equidistant from point A , such that the light signal reaches them both at time t in S . Let a second frame S' translate uniformly with respect to S , in a direction parallel to the line segment joining B and C , such that the material points A, B, C have velocity v with respect to S' . In S' , furthermore, the origin of the primed coordinates x', y', z', t' coincides with that of S at time $t = t' = 0$. A light signal propagating in all directions from point A at time $t_0 = t'_0 = 0$ will reach the surface of a certain sphere at time $t' > 0$, such that:

$$x'^2 + y'^2 + z'^2 - c^2t'^2 = 0. \quad (1.15)$$

The center of the light sphere in S' at time t' coincides with point O , as shown in Figure 1.12, while the origin of S has traveled a distance vt' from O and is located at point A' with respect to frame S' .⁷² With these preliminaries in place, Laue concluded that in frame S' there exists no value of time t' for which the material

Fig. 1.12 Meridional section of a simultaneity light sphere for frame S' with origin O , after Laue (1911a, 35).



⁷¹Planck's argument, which builds on that of Einstein (see above, § 1.2), has inspired many textbook authors. For an example employing a spherical array of photomultipliers at rest in two inertial frames in relative motion, see Rosser (1967, 76).

⁷²Laue's use of primes in his light-sphere diagram is peculiar, but is reproduced intact in Figure 1.12, in keeping with the first four editions of his textbook (up to 1921). In the sixth edition (von Laue 1955, 29), $A, B,$ and C are all unprimed, and the primed symbols are as expected: O' and t' .

points B and C lie on the same spherical surface and that, consequently, B and C are not reached by the light signal simultaneously in frame S' .

What the diagram shows, Laue wrote, is that in frame S' , a light signal reaches point C before it reaches point B . For every inertial system, he concluded, there is a “particular time, differing from that of other systems.” For this insight, Laue credited the “acuity and elevated philosophical sense of Einsteinian ideas” and proceeded to derive the Lorentz transformation, invoking the form invariance of the wave equation, along with linearity and symmetry constraints, and noting the fact that the Lorentz transformation follows just as well from the invariance of the sum of squares

$$x^2 + y^2 + z^2 - c^2t^2. \quad (1.16)$$

The light sphere formed the centerpiece of Laue’s discussion of the Lorentz transformation. As mentioned above, Cunningham and Einstein had employed the light-sphere demonstration in 1907. The fact that Laue preferred to invoke the form invariance of the wave equation in his textbook is of no particular significance, as he, too, went on to employ the light-sphere demonstration (see von Laue 1913a, 110).

Laue’s treatise skillfully combined and repackaged results obtained in the new field of relativity, including the Minkowski diagram and the light-sphere interpretation of the Lorentz transformation and simultaneity relations. The treatise featured the light sphere in a discussion of the foundations of relativistic kinematics and, in the revised and extended second edition of 1913, employed the term “light sphere” in this context (*Lichtkugel*, von Laue 1913b, 36).

Both the relativity of simultaneity and the proof of the Lorentz transformation had previously been demonstrated with a light sphere by Otto Berg (1874–1939), a Privatdozent at the University of Greifswald, in an essay entitled “The relativity principle of electrodynamics.” Taking his cue from Minkowski’s bold claim that the new ideas about time and space in relativity theory had sprung from the ground of experimental physics (Minkowski 1909), Berg, an experimental physicist, prefaced his pamphlet with the opinion that “many philosophers will doubt” such a claim. He then set out to examine the “experimental foundations of the principle [of relativity],” a topic Minkowski had scrupulously avoided. In light of Bucherer’s attack on the utility of the light-sphere hypothesis for explaining experimental results, mentioned above (Section 1.2), Berg’s recourse to the light sphere in this essay comes as a surprise. But as Berg observed, the “clarification of ideas” in relativity theory realized by Einstein and Minkowski had “hardly anything” to do with experiments. In any case, Lorentz’s theory was “just as good” as the newer theories, as far as representing the latest experimental results was concerned (Berg 1910, 357).

Berg’s treatment of the relativity of simultaneity differed little from that of Laue, mentioned above, with one exception: Berg did not illustrate his discussion with a diagram. He presented his light-sphere demonstration of the relativity of simultaneity as a “concrete example” of Einstein’s light postulate and one that later

served his derivation of the Lorentz transformation. Laue must have admired Berg's approach, as he employed it without change, apart from the addition of a graphic illustration (redrawn here as Figure 1.12).

Laue's light-sphere-based demonstration of the relativity of simultaneity had another forerunner in the person of Harry Bateman. Following Cunningham's lead, Bateman (1912, 340) reckoned Einstein's light postulate to be equivalent to admitting the existence of a group of transformations for which (1.16) is covariant. By mid-1910, Bateman (1910b, 624) realized that the Lorentz transformation did not alter the form of tangent-oriented spheres. This insight may be what led him to attribute the origin of the Lorentz transformation not to Voigt, in the manner of Minkowski (1909), but to the French differential geometer Albert Ribaucour, known for his contributions to the geometry of spheres (Ribaucour 1870).

In virtue of his understanding of the relation between the Lorentz transformation and the fifteen-parameter group G_{15} , Bateman went on, in a paper completed in December 1910, to compare the results of Cunningham and himself with those of Poincaré and the leading German relativists:

According to the general principles of group-theory, the quantities and relations which are invariant with regard to the principal group should represent true physical quantities and relations. Some of these invariants for the group G_{15} have been found by Einstein, Poincaré, Minkowski, Planck, Cunningham and the author.⁷³ It is desirable that all the principal invariants and covariants for the group should be found, for then we shall perhaps be able to decide whether Einstein's conditions of observation are the right ones or not. (Bateman 1912, 340)

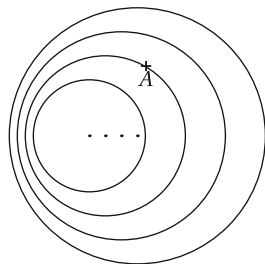
Bateman's accomplishment was duly recognized by Philipp Frank (1884–1966), with whom he probably crossed paths in Göttingen in 1906. Frank (1911) described the covariance of the Maxwell equations under the Lorentz group as “one of the most important mathematico-physical facts of modern physics” and identified Lorentz, Minkowski, and Bateman as the principal investigators in this area of study, to which he and the Viennese mathematician Hermann Rothe (1882–1923) were active contributors, along with von Ignatowsky.

In a wide-ranging review of the consequences of relativity theory for the philosophy of space and time, Bateman drew on Ribaucour's transformations of spheres; his idea was to investigate the “physical aspect of time in order to understand the idea of simultaneity” (Bateman 1910a, 2). In what Bateman called a “view,” an ordered pair of spheres represents a four-vector, the components of which are differences of spatial coordinates and radii. Four-vector magnitude is given by the length of a shared tangent, such that that a null vector corresponds to spheres in contact.

By considering Ribaucour's spheres as light spheres, Bateman demonstrated the relativity of simultaneity and the impossibility of hyperlight velocities. Bateman's

⁷³The transformations of the 15-parameter group of conformal transformations G_{15} correspond to what Bateman called the “spherical wave transformations.” On the Bateman-Cunningham discovery of the covariance of Maxwell's equations under G_{15} , see Rowe (1999, 211), and Kastrop (2008).

Fig. 1.13 A light source in motion, redrawn from Bateman (1910a).



depiction of a light source in uniform translation (Figure 1.13) features four nonconcentric light spheres. His figure differs little from that employed by Poincaré in 1906–1907 (Figure 1.3); only the direction of motion is reversed. Imagining a space filled with light-synchronized clocks, Bateman argued with respect to his diagram that the wave front of only one light sphere may pass through a given point *A* at a given time, such that behavior of light waves makes manifest the simultaneity relation.

In subsequent papers, Bateman neglected to discuss or apply his diagrammatic interpretation of four-vectors, which quickly fell from view. A similar interpretation of four-vectors, proposed by a professor of descriptive geometry in Braunschweig, Timerding (1912, 1915), fared no better. Few theorists in Britain were then familiar with four-vectors, and consequently, few were in a position to grasp the full meaning of Bateman’s potent image of tangent spheres. Bateman’s illustration of the Einstein simultaneity relation was thereby less comprehensible to his contemporaries than the simpler one concocted by Laue, which involved only a passing knowledge of plane geometry.

1.9 Light-Figure Skepticism

More than a few physicists felt that the grounds for accepting light-sphere covariance as the foundation of relativity were not compelling. For example, as mentioned above, A. H. Bucherer saw no need to adopt Einstein’s view of the light sphere. This section takes up the cases of two light-figure skeptics, which is to say, physicists who contested the epistemic priority accorded by Einstein to light waves: Emil Wiechert (1861–1928) and Waldemar von Ignatowsky (1875–1943).

Minkowski’s colleague in Göttingen, the geophysicist Emil Wiechert admired Minkowski’s theory, but like many scientists, he remained attached to the notion of an ether. Wiechert’s ether was attached by stipulation to an inertial frame and was entirely consistent, in his view, with Minkowski’s spacetime theory (Wiechert 1911, 757). In an essay entitled “The principle of relativity and the ether,” Wiechert held that both sound waves and light waves that are spherical in one (absolute) frame are flattened in the direction of motion of an observer translating with respect to this frame. Even if Wiechert granted that Einstein was the first to develop a “rigorous

understanding of Lorentz covariance” and to understand the conventional nature of distant simultaneity, he did not feel compelled to adopt Einstein’s kinematics.⁷⁴ In a review of Wiechert’s theory, Laue (1912) found fault with his logic, but he admitted with Wiechert that the question of the existence of an absolute frame belonged to philosophy, not to physics.

Like Wiechert, von Ignatowsky admired Minkowski’s spacetime theory, but was dissatisfied with Einstein’s relativity. In particular, Einstein’s light postulate seemed unobvious to him. Nonetheless, he was impressed by how the light constant c appeared in Minkowskian relativity to be “more a universal spacetime constant than the speed of light” (von Ignatowsky 1910, 793).⁷⁵ What von Ignatowsky sought to derive were coordinate transformations that guarantee relativity of inertial frames, but do not depend on the light postulate. Introducing the usual constraints, and denoting by p a differential quotient depending on position x , time t , and velocity v , he found (in modified notation):⁷⁶

$$dx' = pdx - pvdt, \quad dt' = -pvndx + pdt, \quad (1.17)$$

where n is a universal constant, such that

$$p^2 = 1/(1 - v^2n). \quad (1.18)$$

In order to determine the value of the constant n , von Ignatowsky considered the equipotential surface of a point charge in uniform motion. For a comoving observer, the equipotential surface has the form of a sphere, but for an observer at rest, the equipotential surface is described by a Heaviside ellipsoid. In other words, for an observer at rest, the equipotential surface is a rotational ellipsoid, the longitudinal and transverse axes of which are related by $1 :: \sqrt{1 - v^2/c^2}$. Von Ignatowsky’s transformation requires that a spherical surface attached to the comoving observer’s frame transform to a flattened ellipsoid in the frame of an observer at rest, such that

$$\sqrt{1 - v^2/c^2} = \sqrt{1 - v^2n}. \quad (1.19)$$

Consequently, $n = 1/c^2$, such that both c and n are now universal constants. Von Ignatowsky’s identification of c as a universal constant depends on the form of the equipotential surface, a form independent of the principle of relativity, as von Ignatowsky was careful to point out.⁷⁷

⁷⁴Wiechert to Lorentz, 9 March 1912, in Kox (2008, 359); Wiechert (1911, 756).

⁷⁵Born in Tiflis (Tbilisi, Georgia), von Ignatowsky earned his Ph.D. in physics at the University of Giessen in 1909 and found employment with the Leitz optical firm in Wetzlar (Klein et al. 1993, 251).

⁷⁶On von Ignatowsky’s transformation see Jammer (1979, 215), Torretti (1996, 76), Brown (2005, 105), and Darrigol (2014, 139).

⁷⁷On the relation between Lorentz contraction and the Heaviside ellipsoid, see Hunt (1988).

Von Ignatowsky noted in passing that v represents the “speed of one of our worlds,” i.e., one of “unendlessly many reference frames.” While the latter description recalls Cunningham’s multiple-ether view of relativity (Section 1.2), it is more likely an additional echo of Minkowski’s Cologne lecture. In a final tribute to Minkowski’s spacetime realism, von Ignatowsky concluded his derivation with the following credo:⁷⁸

Now we should not consider an inertial coordinate system as something like a mere mathematical entity, but we must instead think of it as a material world with its observers and synchronized clocks. (von Ignatowsky 1910, 794)

Although von Ignatowsky drew freely on Minkowskian language and imagery, he did not adopt Minkowski’s four-dimensional calculus, preferring to rewrite the latter’s four-dimensional “vectors of the first and second type” in the form of ordinary three-vectors. More than likely, this was a choice guided by his recent investment in three-dimensional vector analysis, in the form of a book (von Ignatowsky 1909) published in Eugen Jahnke’s Teubner collection “*Mathematisch-Physikalische Schriften für Ingenieure und Studierende*.”

Wiechert and von Ignatowsky were uneasy with the special evidentiary status accorded by Einstein to light signals, as reflected in their approaches to relativity. Wiechert’s rejection of Einstein’s radical reform of kinematics found inspiration in Minkowski’s spacetime theory, in virtue of the absolute nature of the direction of a particle’s four-velocity (Wiechert 1911, 757). Von Ignatowsky’s admiration for Minkowski’s geometric interpretation of the light constant c and concomitant rejection of the latter’s four-dimensional formalism suggest that one could accommodate a Minkowskian ontology while rejecting Minkowskian formalism, which is to say, the precise opposite of what Laue advised in his textbook (Walter 2010).

1.10 Discussion

Introduced without fanfare by Einstein in connection with his postulates of relativity and universal lightspeed invariance, the notion of light-sphere covariance engaged the imagination of theorists and experimentalists alike. A rival light-shell theory was soon proposed, in the form of Poincaré’s light ellipsoid and assorted two-dimensional diagrams. At that time, Poincaré’s theory of the light ellipsoid had much to recommend it, including a privileged coordinate frame and a simple diagram-based derivation of the Lorentz transformation. Poincaré did not seek to publish his derivation, however, and soon Cunningham and Einstein published their own equally elementary algebraic derivations of the Lorentz transformation, based on light-sphere covariance.

⁷⁸“Nun dürfen wir aber unter einem Ruhekoordinatensystem nicht etwa nur ein mathematisches Gebilde verstehen, sonder müssen uns dabei eine materielle Welt mit ihren Beobachtern und synchronem Uhren denken.”

Einstein's kinematics lacked a visually intuitive model until 1908, when Minkowski proposed a model of spacetime that subtended an elementary geometric derivation of the Lorentz transformation, albeit a derivation that Minkowski did not see fit to publish himself. Minkowski did not discuss the light sphere directly, either. Instead, he presented the Lorentz covariance of the light-wave equation as a compelling formal argument in favor of his four-dimensional approach to physics, and cast light-sphere covariance in four-dimensional language, introducing the lightcone structure of spacetime. His expression of the Lorentz transformation as a 4×4 matrix inspired investigations by Cunningham and Bateman of the conformal covariance of Maxwell's equations, which suggested the possibility of a generalization of the principle of relativity to frames in non-inertial motion. The lightcone itself inspired Robb's theory of space and time, intended originally as an alternative to Einsteinian relativity.

The idea of light-sphere covariance traveled across both national and disciplinary boundaries, being carried initially by journals of physics, philosophy, and general science in Germany and Great Britain. When Max Laue wrote his treatise on the principle of relativity (Laue 1911a), he passed over the contributions of Cunningham and Bateman and drew instead on the work of a fellow German Privatdozent, Otto Berg. Cunningham and Bateman were colleagues for a year, while Bateman and Laue heard lectures by Hilbert and Minkowski in Göttingen. All of these scientists were young men; only Berg had passed thirty. None held a permanent university position at the time of the contributions studied here, and all but Berg went on to obtain academic appointments.⁷⁹

According to the recollections of Frank (1947, 206), Einstein remarked to him that he could "hardly understand Laue's book." Frank read Einstein's comment as a reflection on the mathematical sophistication of Laue's treatise, but the mathematics employed by Laue were certainly not new to Einstein. In light of the several contributions to relativity theory and the theory of the electron after 1905 reviewed in this chapter, I suggest an alternative reading of Einstein's offhand comment to Frank on Laue's treatise: Einstein found the content of Laue's book to stem in large part from the work of others; as such, for Einstein Laue's book was not incomprehensible, just foreign to his own way of thinking about relativity.

Although the concept of light-sphere covariance crossed national and disciplinary boundaries with apparent ease, it did not meet with universal assent. There were those, like Bucherer, who found Einstein's argument in favor of light-sphere covariance to be unconvincing. The special evidentiary status assigned to the behavior of light waves in Einstein's theory troubled physicists like von Ignatowsky and Wiechert, who sought to treat electromagnetic waves in the same manner as other propagation phenomena. The case of Poincaré and Robb is particularly

⁷⁹Berg went to work for the Siemens-Halske engineering firm in Berlin, where he co-discovered element 75 (Rhenium) with Walter Noddack and Ida Tacke.

instructive in this regard, in that they both shared Einstein's high epistemic regard for light waves, but deplored – for different philosophical reasons – Einstein's metric interpretation of the light sphere.

Four decades after the events described in this chapter, von Laue wrote about the “somewhat excessive polemic” against relativity as a consequence of “lack of insight” on the part of the theory's opponents (von Laue 1947, 68). The form of a light pulse for moving observers was a topic about which leading theorists disagreed, as we have seen. Von Laue also recalled a “decisive turn” for relativity theory, triggered by Einstein's view of the equal epistemic value of space and time measurements among inertial frames of reference. Closely related to Einstein's belief, the derivation of the Lorentz transformation via covariance of the light-sphere equation stabilized interpretations of the transformation along Einsteinian lines and contributed powerfully to the emergence of a unified doctrine of the physics of inertial frames. One consequence of this movement was a heightened recognition of Einstein as the principal architect of the theory of relativity, as expressed by Laue's 1911 treatise and its six re-editions.

Acknowledgements Key points of this paper were elaborated in discussions with Olivier Darrigol, Tilman Sauer, June Barrow-Green, and David Rowe; their help is much appreciated. The paper benefits from the expert assistance of Kathryn McKee and Fiona Colbert of St. John's College, whom I thank most kindly. Citations of the Joseph Larmor Correspondence are by permission of the Master and Fellows of St. John's College, Cambridge. Permission to quote from the Council Minutes of the London Mathematical Society is gratefully acknowledged. I thank the Niedersächsische Staats- und Universitätsbibliothek Göttingen for authorizing publication of the diagram in Figure 1.9. I am grateful for the support of the Dibner Rare Book and Manuscript Library and to its staff members Lilla Vekerdy and Kirsten van der Veen for their able assistance during my residence in 2013. A preliminary version of the paper was published in 2011 on [PhilSci-Archive](#).

References

- Abraham, M. (1905). *Theorie der Elektrizität, Volume 2, Elektromagnetische Theorie der Strahlung*. Leipzig: Teubner.
- Balázs, N. L. (1972). The acceptability of physical theories: Poincaré versus Einstein. In L. O'Raiifeartaigh (Ed.) *General relativity: Papers in Honour of J. L. Synge* (pp. 21–34). Oxford: Oxford University Press.
- Barrow-Green, J., & Gray, J. (2006). Geometry at Cambridge, 1863–1940. *Historia Mathematica*, 33(3), 315–356.
- Bateman, H. (1908). On conformal transformations of a space of four dimensions and their application to geometrical optics. *Report-British Association*, 78, 627–629.
- Bateman, H. (1909). The conformal transformations of a space of four dimensions and their applications to geometrical optics. *Proceedings of the London Mathematical Society*, 7, 70–89.
- Bateman, H. (1910a). The physical aspect of time. *Memoirs and Proceedings, Manchester Literary and Philosophical Society*, 54(14), 1–13.
- Bateman, H. (1910b). The relation between electromagnetism and geometry. *Philosophical Magazine*, 20, 623–628.

- Bateman, H. (1910c). The transformations of the electrodynamical equations. *Proceedings of the London Mathematical Society*, 8, 223–264.
- Bateman, H. (1912). Some geometrical theorems connected with Laplace's equation and the equation of wave motion. *American Journal of Mathematics*, 34, 325–360.
- Berg, O. (1910). Das Relativitätsprinzip der Elektrodynamik. *Abhandlungen der Fries'schen Schule*, 3, 333–382.
- Bergson, H. (1922). *Durée et simultanéité : à propos de la théorie d'Einstein*. Paris: Alcan.
- Blumenthal, O. (Ed.). (1913). *Das Relativitätsprinzip; Eine Sammlung von Abhandlungen*. Leipzig: Teubner.
- Born, M. (1909). Die Theorie des starren Elektrons in der Kinematik des Relativitätsprinzips. *Annalen der Physik*, 335, 1–56.
- Born, M. (1912). Besprechung von Max Laue, Das Relativitätsprinzip. *Physikalische Zeitschrift*, 13, 175–176.
- Born, M. (1914). Besprechung von Max Weinstein, Die Physik der bewegten Materie und die Relativitätstheorie. *Physikalische Zeitschrift*, 15, 676.
- Born, M. (1920). *Die Relativitätstheorie Einsteins und ihre physikalischen Grundlagen*. Berlin: Springer.
- Born, M. (1959). Erinnerungen an Hermann Minkowski zur 50. Wiederkehr seines Todestages. *Naturwissenschaften*, 46(17), 501–505.
- Bridgman, P. W. (1916). Tolman's principle of similitude. *Physical Review*, 8, 423–431.
- Briginschaw, A. J. (1979). The axiomatic geometry of space-time: An assessment of the work of A. A. Robb. *Centaurus*, 22(4), 315–323.
- Bromwich, T. J. I. (1901). Conformal space transformations. *Proceedings of the London Mathematical Society*, 33(749), 185–192.
- Brown, H. R. (2005). *Physical relativity: Space-time structure from a dynamical perspective*. Oxford: Oxford University Press.
- Bucherer, A. H. (1907). On a new principle of relativity in electromagnetism. *Philosophical Magazine*, 13, 413–420.
- Bucherer, A. H. (1908a). Messungen an Becquerelstrahlen; die experimentelle Bestätigung der Lorentz-Einsteinschen Theorie. *Physikalische Zeitschrift*, 9, 755–762.
- Bucherer, A. H. (1908b). On the principle of relativity and on the electromagnetic mass of the electron; a reply to Mr. E. Cunningham. *Philosophical Magazine*, 15, 316–318.
- Buchwald, J. Z. (1985). *From Maxwell to microphysics*. Chicago: University of Chicago Press.
- Cat, J. (2016). Images and logic of the light cone: Tracking Robb's postulational turn in physical geometry. *Revista de Humanidades de Valparaíso*, 4(8), 43–105.
- Cunningham, E. (1907). On the electromagnetic mass of a moving electron. *Philosophical Magazine*, 14:538–547.
- Cunningham, E. (1910). The principle of relativity in electrodynamics and an extension thereof. *Proceedings of the London Mathematical Society*, 8, 77–98.
- Cunningham, E. (1911). The principle of relativity. *Report-British Association*, 81:236–245.
- Cunningham, E. (1914). *The principle of relativity*. Cambridge: Cambridge University Press.
- Cuvaj, C. (1970). *A history of relativity: The role of Henri Poincaré and Paul Langevin*. PhD thesis, Yeshiva University, New York.
- Darrigol, O. (1995). Henri Poincaré's criticism of fin de siècle electrodynamics. *Studies in History and Philosophy of Modern Physics*, 26, 1–44.
- Darrigol, O. (1996). The electrodynamic origins of relativity theory. *Historical Studies in the Physical and Biological Sciences*, 26(2), 241–312.
- Darrigol, O. (2000). *Electrodynamics from Ampère to Einstein*. Oxford: Oxford University Press.
- Darrigol, O. (2014). *Physics and necessity: Rationalist pursuits from the Cartesian past to the quantum present*. Oxford: Oxford University Press.
- Darrigol, O. (2015). Poincaré and light. In B. Duplantier & V. Rivasseau (Eds.), *Henri Poincaré, 1912–2012. Progress in mathematical physics* (Vol. 67, pp. 1–50). Berlin: Springer.
- Du Pasquier, L.-G. (1922). *Le principe de la relativité et les théories d'Einstein*. Paris: Doin.
- Einstein, A. (1905). Zur Elektrodynamik bewegter Körper. *Annalen der Physik*, 322, 891–921.

- Einstein, A. (1907). Relativitätsprinzip und die aus demselben gezogenen Folgerungen. *Jahrbuch der Radioaktivität und Elektronik*, 4, 411–462.
- Frank, P. G. (1911). Das Verhalten der elektromagnetischen Feldgleichungen gegenüber linearen Transformationen. *Annalen der Physik*, 340, 599–607.
- Frank, P. G. (1947). *Einstein: His life and times*. New York: Knopf.
- Galison, P. (1979). Minkowski's spacetime: From visual thinking to the absolute world. *Historical Studies in the Physical Sciences*, 10, 85–121.
- Galison, P. (2003). *Einstein's clocks and Poincaré's maps: Empires of time*. New York: Norton.
- Goldberg, S. (1970). In defense of Ether: The British response to Einstein's special theory of relativity 1905–1911. *Historical Studies in the Physical Sciences*, 2, 89–125.
- Gray, J. (1989). *Ideas of space: Euclidean, non-Euclidean and relativistic* (2nd ed.). Oxford: Clarendon.
- Gray, J. (2013). *Henri Poincaré: A scientific biography*. Princeton: Princeton University Press.
- Guillaume, E. (1921). Expression mono et polyparamétrique du temps dans la théorie de la relativité. In H. Villat (Ed.), *Comptes rendus du Congrès international des mathématiciens: Strasbourg, 22–30 septembre 1920* (pp. 594–602). Toulouse: Edouard Privat.
- Guillaume, E. (1922). Un résultat des discussions de la théorie de la relativité d'Einstein au Collège de France. *Revue générale des sciences pures et appliquées*, 33, 322–324.
- Heilbron, J. L. (1963). Interview with Dr. Ebenezer Cunningham. <http://www.aip.org/history/catalog/icos/4565.html>
- Heilbron, J. L. (1986). *The dilemmas of an upright man: Max Planck as spokesman for German science*. Berkeley: University of California Press.
- Hirosgie, T. (1968). Theory of relativity and the ether. *Japanese Studies in the History of Science*, 7, 37–53.
- Holton, G. (1965). The metaphor of space-time events in science. *Eranos Jahrbuch*, 34, 33–78.
- Hunt, B. J. (1986). Experimenting on the ether: Oliver J. Lodge and the great whirling machine. *Historical Studies in the Physical Sciences*, 16, 111–134.
- Hunt, B. J. (1988). The origins of the FitzGerald contraction. *British Journal for the History of Science*, 21(1), 67–76.
- Jammer, M. (1979). Some foundational problems in the special theory of relativity. In G. Toraldo di Francia (Ed.), *Problems in the foundations of physics. Proceedings of the international school of physics "Enrico Fermi"* (Vol. 72, pp. 202–236). Amsterdam: North-Holland.
- Janssen, M., & Mecklenburg, M. (2006). From classical to relativistic mechanics: Electromagnetic models of the electron. In V.F. Hendricks, K.F. Jørgensen, J. Lützen, & S.A. Pedersen (Eds.), *Interactions: Mathematics, physics and philosophy, 1860–1930* (pp. 65–134). Dordrecht: Springer.
- Jungnickel, C., & McCormach, R. (1986). *Intellectual mastery of nature: Theoretical physics from Ohm to Einstein*. Chicago: University of Chicago Press.
- Kastrup, H. A. (2008). On the advancements of conformal transformations and their associated symmetries in geometry and theoretical physics. *Annalen der Physik*, 17(9), 691–704.
- Kennedy, W. L. (2005). On Einstein's 1905 electrodynamics paper. *Studies in History and Philosophy of Modern Physics*, 36(1), 61–65.
- Klein, M. J., Kox, A. J., & Schulmann, R. (Eds.). (1993). *The collected papers of Albert Einstein, volume 5, the Swiss years: Correspondence, 1902–1914*. Princeton: Princeton University Press.
- Kormos Buchwald, D., Sauer, T., Rosenkranz, Z., Illy, J., & Holmes, V. I. (Eds.). (2006). *The collected papers of Albert Einstein, volume 10, the Berlin years: Correspondence, May–December 1920*. Princeton: Princeton University Press.
- Kox, A. J. (Ed). (2008). *The scientific correspondence of H.A. Lorentz* (Vol. 1). Berlin: Springer.
- Langevin, P. (1905). Sur l'origine des radiations et l'inertie électromagnétique. *Journal de physique théorique et appliquée*, 4, 165–183.
- Langevin, P. (1906). The relations of the physics of electrons to the other branches of science. In H. J. Rogers (Ed.), *Congress of arts and science, universal exposition, St. Louis, 1904, volume 4: Physics, chemistry, astronomy, sciences of the Earth*, pages 121–156. Houghton, Mifflin & Co., Boston.

- Larmor, J. (1900). *Aether and matter*. Cambridge: Cambridge University Press.
- Larmor, J. (1938). Alfred Arthur Robb 1873–1936. *Obituary Notices of Fellows of the Royal Society*, 2, 315–321.
- Laue, M. (1903). *Über die Interferenzerscheinungen an planparallelen Platten*. PhD thesis, University of Berlin, Berlin.
- Laue, M. (1907). Die Mitführung des Lichtes durch bewegte Körper nach dem Relativitätsprinzip. *Annalen der Physik*, 328, 989–990.
- Laue, M. (1908). Die Wellenstrahlung einer bewegten Punktladung nach dem Relativitätsprinzip. *Berichte der Deutschen Physikalischen Gesellschaft*, 10, 838–844.
- Laue, M. (1911a). *Das Relativitätsprinzip*. Braunschweig: Vieweg.
- Laue, M. (1911b). Zur Dynamik der Relativitätstheorie. *Annalen der Physik*, 340, 524–542.
- Laue, M. (1912). Zwei Einwände gegen die Relativitätstheorie und ihre Widerlegung. *Physikalische Zeitschrift*, 13, 118–120.
- Lémeray, M. (1912). Sur un théorème de M. Einstein. *Comptes rendus hebdomadaires de l'Académie des sciences de Paris*, 155, 1224–1227.
- Lie, S., & Scheffers, G. (1893). *Vorlesungen über kontinuierliche Gruppen mit geometrischen und anderen Anwendungen*. Leipzig: Teubner.
- Liénard, A. (1898). Champ électrique et magnétique produit par une charge concentrée en un point et animée d'un mouvement quelconque. *Éclairage électrique*, 16, 5–14, 53–59, 106–112.
- Lorentz, H. A. (1897). Ueber den Einfluss magnetischer Kräfte auf die Emission des Lichtes. *Annalen der Physik*, 299, 278–284.
- Lorentz, H. A. (1904). Electromagnetic phenomena in a system moving with any velocity less than that of light. *Proceedings of the Section of Sciences, Koninklijke Akademie van Wetenschappen te Amsterdam*, 6, 809–831.
- Lorentz, H. A. (1909). *The theory of electrons and its application to the phenomena of light and radiant heat*. New York: Columbia University Press.
- Lorentz, H. A. (1916). *The theory of electrons and its application to the phenomena of light and radiant heat* (2nd ed.). Leipzig: Teubner.
- Lorenz, L. V. (1867). On the identity of the vibrations of light with electrical currents. *Philosophical Magazine*, 34(230), 287–301.
- Maltese, G., & Orlando, L. (1995). The definition of rigidity in the special theory of relativity and the genesis of the general theory of relativity. *Studies in History and Philosophy of Modern Physics*, 26(3), 263–306.
- Manegold, K.-H. (1970). *Universität, Technische Hochschule und Industrie*. Berlin: Duncker & Humblot.
- Martínez, A. A. (2009). *Kinematics: The lost origins of Einstein's relativity*. Baltimore: Johns Hopkins University Press.
- McCrea, W. H. (1978). Ebenezer Cunningham. *Bulletin of the London Mathematical Society*, 10(1), 116–126.
- Miller, A. I. (1973). A study of Henri Poincaré's 'Sur la dynamique de l'électron'. *Archive for History of Exact Sciences*, 10, 207–328.
- Miller, A. I. (1976). On Einstein, light quanta, radiation and relativity in 1905. *American Journal of Physics*, 44(10), 912–923.
- Miller, A. I. (1981). *Albert Einstein's special theory of relativity: Emergence (1905) and early interpretation*. Reading, MA: Addison-Wesley.
- Minkowski, H. (1908). Die Grundgleichungen für die electromagnetischen Vorgänge in bewegten Körpern. *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen*, 1908, 53–111.
- Minkowski, H. (1909). Raum und Zeit. *Jahresbericht der deutschen Mathematiker-Vereinigung*, 18, 75–88.
- Minkowski, H. (1915). Das Relativitätsprinzip. *Jahresbericht der deutschen Mathematiker-Vereinigung*, 24, 372–382.
- Neumann, C. G. (1870). *Ueber die Principien der Galilei-Newton'schen Theorie*. Leipzig: Teubner.

- Newman, E. T., & Price, R. H. (2010). The Lorentz transformation: Simplification through complexification. *American Journal of Physics*, 78(1), 14–19.
- Norton, J. D. (1992). Einstein, Nordström and the early demise of Lorentz-covariant theories of gravitation. *Archive for History of Exact Sciences*, 45, 17–94.
- Penrose, R. (1959). The apparent shape of a relativistically moving sphere. *Proceedings of the Cambridge Philosophical Society*, 55, 137–139.
- Planck, M. (1907). Zur Dynamik bewegter Systeme. *Sitzungsberichte der königliche preußischen Akademie der Wissenschaften*, 29, 542–570.
- Planck, M. (1910). *Acht Vorlesungen über theoretische Physik*. Leipzig: Hirzel.
- Poincaré, H. (1891). Sur la théorie des oscillations hertziennes. *Comptes rendus hebdomadaires de l'Académie des sciences de Paris*, 113, 515–519.
- Poincaré, H. (1901). *Électricité et optique: la lumière et les théories électrodynamiques*. Paris: Carré et Naud.
- Poincaré, H. (1904). L'état actuel et l'avenir de la physique mathématique. *Bulletin des sciences mathématiques*, 28, 302–324.
- Poincaré, H. (1905a). *Science and hypothesis*. London: Walter Scott.
- Poincaré, H. (1905b). Sur la dynamique de l'électron. *Comptes rendus hebdomadaires de l'Académie des sciences de Paris*, 140, 1504–1508.
- Poincaré, H. (1906). Sur la dynamique de l'électron. *Rendiconti del circolo matematico di Palermo*, 21, 129–176.
- Poincaré, H. (1907). La relativité de l'espace. *Année psychologique*, 13, 1–17.
- Poincaré, H. (1908a). La dynamique de l'électron. *Revue générale des sciences pures et appliquées*, 19, 386–402.
- Poincaré, H. (1908b). *Science et méthode*. Paris: Flammarion.
- Poincaré, H. (1909). La mécanique nouvelle. *Revue scientifique*, 12, 170–177.
- Poincaré, H. (1913). *La dynamique de l'électron*. Paris: Dumas.
- Poincaré, H. (1953). Les limites de la loi de Newton. *Bulletin astronomique*, 17(3), 21–269.
- Pyenson, L. (1979). Physics in the shadow of mathematics: The Göttingen electron-theory seminar of 1905. *Archive for History of Exact Sciences*, 21(1), 55–89.
- Renn, J. (2007). The third way to general relativity. In J. Renn (Ed.), *The genesis of general relativity, volume 3, gravitation in the twilight of classical physics: Between mechanics, field theory, and astronomy. The genesis of general relativity* (pages 21–75). Berlin: Springer.
- Reynolds, W. F. (1993). Hyperbolic geometry on a hyperboloid. *American Mathematical Monthly*, 100, 442–455.
- Ribaucour, A. (1870). Sur la déformation des surfaces. *Comptes rendus hebdomadaires de l'Académie des sciences de Paris*, 70, 330–333.
- Robb, A. A. (1904). *Beiträge zur Theorie des Zeemaneffektes*. PhD thesis, Georg-Augusta Universität Göttingen, Leipzig.
- Robb, A. A. (1911). *Optical geometry of motion: A new view of the theory of relativity*. Cambridge: W. Heffer and Sons.
- Rohrlich, F. (2007). *Classical charged particles* (3rd ed.). Singapore: World Scientific.
- Rosser, W. G. V. (1967). *Introductory relativity*. London: Butterworth.
- Rowe, D. E. (1989). Klein, Hilbert, and the Göttingen mathematical tradition. In K. M. Olesko (Ed.), *Science in Germany. Osiris* (Vol. 5, pp. 186–213). Philadelphia: History of Science Society.
- Rowe, D. E. (1999). The Göttingen response to general relativity and Emmy Noether's theorems. In J. Gray (Ed.), *The symbolic universe: Geometry and physics, 1890–1930* (pp. 189–233). Oxford: Oxford University Press.
- Rowe, D. E. (2008). Max von Laue's role in the relativity revolution. *Mathematical Intelligencer*, 30(3), 54–60.
- Rowe, D. E. (2009). A look back at Hermann Minkowski's Cologne lecture 'Raum und Zeit'. *Mathematical Intelligencer*, 31(2), 27–39.

- Sánchez-Ron, J. M. (1987). The reception of special relativity in Great Britain. In T. F. Glick (Ed.), *The comparative reception of relativity. Boston studies in the philosophy of science* (pp. 27–58). Dordrecht: Reidel.
- Searle, G. F. C. (1897). On the steady motion of an electrified ellipsoid. *Philosophical Magazine*, 44, 329–341.
- Seelig, C. (1960). *Albert Einstein: Leben und Werk eines Genies unserer Zeit*. Zürich: Europa Verlag.
- Silberstein, L. (1914). *The theory of relativity*. London: Macmillan.
- Sommerfeld, A. (1913). Anmerkungen zu Minkowski, Raum und Zeit. In O. Blumenthal (Ed.), *Das Relativitätsprinzip; Eine Sammlung von Abhandlungen* (pp. 69–73). Leipzig: Teubner.
- Sommerfeld, A. (1919). *Atombau und Spektrallinien*. Braunschweig: Vieweg.
- Sommerfeld, A. (1948). *Vorlesungen über theoretische Physik, Bd. 3: Elektrodynamik*. Wiesbaden: Dieterich.
- Stachel, J., Cassidy, D. C., Renn, J., & Schulmann, R. (Eds.). (1989). *The collected papers of Albert Einstein, volume 2, the Swiss years: Writings, 1900–1909*. Princeton: Princeton University Press.
- Staley, R. (1998). On the histories of relativity: The propagation and elaboration of relativity theory in participant histories in Germany, 1905–1911. *Isis*, 89(2), 263–299.
- Sternberg, S. (1986). Imagery in scientific thought by Arthur I. Miller. *Mathematical Intelligencer*, 8(2), 65–74.
- Tanner, J. R. (Ed.). (1917). *The historical register of the University of Cambridge, being a supplement to the calendar with a record of university offices, honours and distinctions to the year 1910*. Cambridge: Cambridge University Press.
- Thomson, J. J. (1893). *Notes on recent researches in electricity and magnetism: Intended as a sequel to Professor Clerk-Maxwell's treatise on electricity and magnetism*. Oxford: Clarendon.
- Timerding, H. E. (1912). Über ein einfaches geometrisches Bild der Raumzeitwelt Minkowskis. *Jahresbericht der deutschen Mathematiker-Vereinigung*, 21, 274–285.
- Timerding, H. E. (1915). Über die Raumzeitvektoren und ihre geometrische Behandlung. In K. Bergwitz (Ed.), *Festschrift Julius Elster und Hans Geitel zum sechzigsten Geburtstage: Arbeiten aus den Gebieten der Physik, Mathematik, Chemie* (pp. 514–520). Braunschweig: Vieweg.
- Tolman, R. C. (1910). The second postulate of relativity. *Physical Review* 31, 26–40.
- Tolman, R. C. (1914). The principle of similitude. *Physical Review*, 3, 244–255.
- Torretti, R. (1996). *Relativity and geometry* (2nd ed.). New York: Dover Publications.
- von Ignatowsky, W. S. (1909). *Die Vektoranalysis und Anwendung in der theoretischen Physik*. Leipzig: Teubner.
- von Ignatowsky, W. S. (1910). Einige allgemeine Bemerkungen zum Relativitätsprinzip. *Berichte der Deutschen Physikalischen Gesellschaft*, 12(20), 788–796.
- von Laue, M. (1913a). Das Relativitätsprinzip. *Jahrbücher der Philosophie*, 1, 99–128.
- von Laue, M. (1913b). *Das Relativitätsprinzip* (2nd ed.). Braunschweig: Vieweg.
- von Laue, M. (1947). *Geschichte der Physik*. Bonn: Universitäts-Verlag.
- von Laue, M. (1952). Mein physikalischer Werdegang; eine Selbstdarstellung. In H. Hartmann (Ed.), *Schöpfer des neuen Weltbildes: Große Physiker unserer Zeit* (pp. 178–207). Bonn: Athenäum-Verlag.
- von Laue, M. (1955). *Die Relativitätstheorie: die spezielle Relativitätstheorie* (6th ed.). Braunschweig: Vieweg.
- von Laue, M. (1961). *Aufsätze und Vorträge*. Braunschweig: Vieweg.
- Walter, S. A. (1999a). Minkowski, mathematicians, and the mathematical theory of relativity. In H. Goenner, J. Renn, T. Sauer, & J. Ritter (Eds.), *The expanding worlds of general relativity, Einstein studies* (Vol. 7, pp. 45–86). Boston: Birkhäuser.
- Walter, S. A. (1999b). The non-Euclidean style of Minkowskian relativity. In J. Gray (Ed.), *The symbolic universe: Geometry and physics, 1890–1930* (pp. 91–127). Oxford: Oxford University Press.

- Walter, S. A. (2007). Breaking in the 4-vectors: The four-dimensional movement in gravitation, 1905–1910. In J. Renn & M. Schemmel (Eds.), *The genesis of general relativity, volume 3: Theories of gravitation in the twilight of classical physics* (pp. 193–252). Berlin: Springer.
- Walter, S. A. (2008). Hermann Minkowski's approach to physics. *Mathematische Semesterberichte*, 55(2), 213–235.
- Walter, S. A. (2009). Hypothesis and convention in Poincaré's defense of Galilei spacetime. In M. Heidelberger & G. Schieman (Eds.), *The significance of the hypothetical in the natural sciences* (pp. 193–219). Berlin: Walter de Gruyter.
- Walter, S. A. (2010). Minkowski's modern world. In V. Petkov (Ed.), *Minkowski spacetime: A hundred years later. Fundamental theories of physics* (Vol. 165, pp. 43–61). Berlin: Springer.
- Walter, S. A. (2011). Henri Poincaré, theoretical physics and relativity theory in Paris. In K.-H. Schlote & M. Schneider (Eds.), *Mathematics meets physics: A contribution to their interaction in the 19th and the first half of the 20th century* (pp. 213–239). Frankfurt am Main: Harri Deutsch.
- Walter, S. A. (2014). Poincaré on clocks in motion. *Studies in History and Philosophy of Modern Physics*, 47(1), 131–141.
- Walter, S. A. (2017). Henri Poincaré's life, science, and life in science. *Historia Mathematica*, 44(4), 425–435.
- Walter, S. A. (2018). Poincaré-week in Göttingen, in light of the Hilbert-Poincaré correspondence of 1908–1909. In M. T. Borgato, E. Neuenschwander & I. Passeron (Eds.), *Mathematical correspondences and critical editions* (pp. 189–202). Basel: Birkhäuser.
- Walter, S. A., Bolmont, E., & Coret, A. (Eds.). (2007b). *La correspondance d'Henri Poincaré, volume 2: La correspondance entre Henri Poincaré et les physiciens, chimistes et ingénieurs*. Basel: Birkhäuser.
- Walter, S. A., Nabonmand, P., & Rollet, L. (Eds.). (2016). *Henri Poincaré papers*. <http://henripoincarepapers.univ-nantes.fr>.
- Warwick, A. C. (2003). *Masters of theory: Cambridge and the rise of mathematical physics*. Chicago: University of Chicago.
- Weyl, H. (1922). Das Raumproblem. *Jahresbericht der deutschen Mathematiker-Vereinigung*, 31, 205–221.
- Whittaker, E. T. (1951). *A history of the theories of aether and electricity*. London: T. Nelson.
- Wiechert, E. (1900). Elektrodynamische Elementargesetze. In J. Bosscha (Ed.), *Recueil de travaux offerts par les auteurs à H. A. Lorentz. Archives néerlandaises des sciences exactes et naturelles* (Vol. 5, pp. 549–573). The Hague: Nijhoff.
- Wiechert, E. (1911). Relativitätsprinzip und Äther. *Physikalische Zeitschrift*, 12, 689–707, 737–758.
- Williamson, R. B. (1977). Logical economy in Einstein's 'on the electrodynamics of moving bodies'. *Studies in History and Philosophy of Science*, 8(1), 46–60.
- Wright, S. P. (1975). *Henri Poincaré: A developmental study of his philosophical and scientific thought*. PhD thesis, Harvard University, Cambridge, MA.
- Zahar, E. (1997). Poincaré's philosophy of geometry, or does geometric conventionalism deserve its name? *Studies in History and Philosophy of Modern Physics*, 28, 183–218.

Chapter 2

The Behaviour of Rods and Clocks in General Relativity and the Meaning of the Metric Field



Harvey R. Brown

In this view what general relativity really succeeded in doing was to eliminate geometry from physics. James L. Anderson, 1999

2.1 Introduction

Why is the $g_{\mu\nu}$ field commonly assigned *geometrical* significance in general relativity theory? Why is it often regarded the fabric of space-time itself, or the “arena” of dynamical processes? The standard explanation suggests that in part, this is because rods and clocks survey it (if only approximately) in a way that does not depend on their constitution. Here is how Clifford Will put the point in 2001:

The property that all non-gravitational fields should couple in the same manner to a single gravitational field is sometimes called “universal coupling”. Because of it, one can discuss the metric as a property of space-time itself rather than as a field over space-time. This is because its properties may be measured and studied using a variety of different experimental devices, composed of different non-gravitational fields and particles, and . . . the results will be independent of the device. Thus, for instance, the proper time between two events is characteristic of spacetime and of the location of the events, not of the clocks used to measure it. (Will 2001)

Is this reasoning cogent? The purpose of this brief paper is to raise some doubts that complement those raised by Anderson (2007). In Section 2.2, it is questioned

A preliminary version of this essay with the same title appeared on [arXiv:gr-qc/0911.4440v1](https://arxiv.org/abs/gr-qc/0911.4440v1).

H. R. Brown (✉)

Faculty of Philosophy, University of Oxford, Radcliffe Observatory Quarter 555,
Woodstock Road, Oxford OX2 6GG, UK

e-mail: harvey.brown@philosophy.ox.ac.uk

© Springer Science+Business Media, LLC, part of Springer Nature 2018

D. E. Rowe et al. (eds.), *Beyond Einstein*, Einstein Studies 14,

https://doi.org/10.1007/978-1-4939-7708-6_2

whether the universality of the behaviour of rods and clocks is indeed a basic feature of the theory, with special reference to the case of accelerating clocks. In Section 2.3, a detour is taken through special relativity, in which the changing nature of Einstein's views about the explanation of length contraction and time dilation is brought out by looking at a curious passage in his 1949 *Autobiographical Notes* (Einstein 1969). The notion of "measurement" as it pertains to the behaviour of rods and clocks is then critically analysed. In the final Section 2.4, attention turns back to general relativity. It is stressed that the universal behaviour of rods and clocks, to the extent that it exists, is not a consequence of the form of the Einstein field equations, any more than the very signature of the metric is, or any more than the local validity of special relativity is. The notion that the metric field can be viewed as "a property of space-time itself rather than as a field over space-time" is based on a feature of the theory—related to what Will refers to as universal coupling—that arguably sits some distance away from the central dynamical tenets of the theory. And the notion may even be misleading.

2.2 Clocks and Their Complications

In the last sentence of the above quotation, Will does not explicitly indicate that he is talking about infinitesimally close events. Indeed, it seems natural to assume that his remark extends to the notion of proper time between any two events that are connected by a finite curve that is everywhere time-like, and which represents the world-line of a clock. Of course in this case the proper time is defined relative to the curve. But in so far as the time read by the clock is related in the usual way to the length of the curve defined by the metric (i.e. the integral of ds along the curve, for the line element $ds^2 = g_{\mu\nu}dx^\mu dx^\nu$), it is taken to be a universal phenomenon, i.e., independent on the constitution of the clock.

For most purposes, physical clocks are designated "good" or "ideal" when they tick in step with the temporal parameter appearing in the fundamental equations associated with our best theories of the non-gravitational interactions. But such considerations usually involve freely moving clocks, and even then no clock smaller than the whole universe can act strictly ideally because the action of the rest of the universe on it cannot be entirely screened off. Furthermore, if a very good approximation to an ideal clock can be found in regions of space-time with weak gravitational fields, when the same clock is placed in a strong gravitational field this behaviour will generally not persist, as Anderson (1999) stressed. How strong the tidal forces have to be to cause a significant degree of disruption to the workings of the clock will of course generally depend on the constitution of the clock.

And then there is the matter of accelerating clocks, corresponding to non-geodesic time-like worldlines. For a correlation of the kind mentioned above, between the reading of an accelerating clock and the length of its world-line, to hold, it must be the case that the effect of motion on the clock at an instant can only be related to its instantaneous velocity, and not to its acceleration. This condition

is commonly (and somewhat misleadingly) referred to as the *clock hypothesis* in both special and general relativity.¹ The question to be addressed in the rest of this section is whether the clock hypothesis, to the extent that is satisfied by real clocks, is indeed a universal phenomenon in the above sense. To this end, let's consider first one of the great early experimental tests of general relativity.

The Pound-Rebka experiment (Pound and Rebka 1960a), involving the emission and absorption of gamma rays by Fe^{57} nuclei—using the newly-discovered Mössbauer effect—placed at different heights in the Earth's gravitational field, is widely known for its role in corroborating Einstein's prediction of the general-relativistic red-shift effect.² It is perhaps less widely appreciated that the experiment also has a bearing on time dilation. Both the emitter and absorber nuclei undergo accelerations due to thermal lattice vibrations, and Pound and Rebka (1960b), and independently Josephson (1960), had realized that even a small temperature difference between the emitter and absorber would result in an observable shift in the absorption line as detected in the Doppler shift method.³ This shift, which must be taken into account in testing for the red-shift effect, is a consequence of the differential relativistic time dilation associated with the different root mean square (rms) velocities of the emitter and absorber nuclei. In fact, as was clarified by Sherwin (1960), the Fe^{57} nuclei are playing the role of clocks in the "twin paradox", or "clock retardation" effect. Note that the accelerations involved were of the order of $10^{16}g$, and the accuracy of the experiment was within 10%.⁴

¹For a discussion of the role of the clock hypothesis within special and general relativity, see Brown (2005, section 6.2.1).

²It is noteworthy that this famous Harvard experiment was not the first to test the red-shift hypothesis; it was preceded in 1960 by a similar experiment using the Mössbauer effect performed at Harwell in the UK (Cranshaw et al. 1960). The Harwell group also performed in the same year another red-shift test, again using the Mössbauer effect, but this time involving a source at the centre of a rotating wheel which contained a thin iron absorber ((Hay et al. 1960), see also Sherwin (1960)). For a detailed account of the history of the early redshift experiments, see Hentschel (1996).

In 1960 it was clarified by Schild (1960) that red-shift tests are by nature insensitive to the precise form of Einstein's field equations. In the same paper, he also noted that nonetheless the red-shift phenomenon itself leads naturally to the idea that gravitational fields are related to the curvature of space-time. It is worth emphasizing that this argument for curvature depends on an appeal to the inhomogeneity of the gravitational field and therefore to a series of red-shift experiments sufficiently separated on the surface of the Earth; a single example of the red-shift, such as in the Pound-Rebka experiment, in which the equipment is largely insensitive to tidal forces, is not enough. This point is sometimes overlooked, as in Carroll's otherwise excellent (Carroll 2004, pp. 53–4). For more details, see Brown and Read (2016).

³If, for example, the source is at a higher temperature than the absorber, the shift is negative. For resonance absorption to occur in this case, the absorber must be given a small velocity away from the source so that the Doppler effect can compensate for the shift; see Sherwin (1960, p. 19).

⁴An earlier 1960 redshift test, also involving the Mössbauer effect, involved a source at the centre of a rotating wheel which contained a thin iron absorber ((Hay et al. 1960); see also Sherwin (1960)). Again a frequency shift due to relativistic time dilation—a case of clock retardation—was detected to an accuracy of a few percent; the radial acceleration in this experiment was of the order of 10^4g .

A later and more celebrated version of the retardation experiment was the 1977 CERN $g-2$ experiment ((Bailey et al. 1977); see also Wilkie (1977)) with orbiting muons in a magnetic field and suffering accelerations of the order $10^{18}g$. A relativistic clock retardation effect was reported—the clocks of course being the unstable muons—to within an accuracy of 0.1%.

In all such experiments, the clock retardation is calculated in conformity with the clock hypothesis, so that the effect is due to rms velocities or integration over the instantaneous velocities of the clocks: the instantaneous accelerations themselves are supposed to contribute nothing to the effect. But for any given clock, no matter how ideal its behaviour when moving inertially, there will in principle be an acceleration such that to achieve it the external force acting on the clock will disrupt its inner workings. As Eddington succinctly put it: “We may force it into its track by continually hitting it, but that may not be good for its time-keeping qualities” (Eddington 1966). We can infer from the above experiments that Fe^{57} nuclei and muons *do* satisfy the clock hypothesis under accelerations of at least $10^{16}g$ and $10^{18}g$, respectively. But this happy circumstance depends not on luck, nor definition, but on the physical make-up of the clocks in question.

In the case of the Pound-Rebka experiment, Sherwin provided a back-of-the-envelope calculation of the deformation of the Fe^{57} nucleus caused by accelerations of $10^{16}g$. Given the nature of the gamma ray resonance, the force associated with the relative displacement of the protons and neutrons within the nucleus is 3×10^{28} dynes/cm. And given that only the protons are affected by the electric force,

... the neutron should suffer a maximum relative displacement of about 1 part in 10^{13} of the nuclear diameter. Even using the great sensitivity of the Mössbauer resonance, such a small distortion is not likely to produce an observable effect. First, it would have to produce a relative shift of the same order of magnitude between the two states which define the resonance. Second, the change in the resonance frequency arising from the acceleration would have to be independent of the direction of the acceleration, for, if it were not, the rapidly varying, cyclical acceleration patterns would have their effects averaged to zero over the emission time of a quantum (in a manner similar to that of the first-order Doppler shift arising from lattice vibrations). We conclude from this rough calculation that the mechanical distortion of the nuclear structure under the accelerations due to the lattice vibrations is very small, but under favorable circumstances an intrinsic acceleration-dependent effect in the resonance frequency might be observable. (Sherwin 1960)

In the case of the $g - 2$ muon experiment at CERN, Eisele (1987) provided a detailed analysis of the decay process for muons experiencing centripetal acceleration. With respect to inertial frames, such orbits are described in terms of a Landau-level with high quantum number, and Eisele used perturbation techniques in the theory of the weak interaction to calculate the approximate life time of the muons. He concluded that the correction to the calculation based on the clock hypothesis for accelerations of $10^{18}g$ would be less than 1 part in 10^{25} , many orders of magnitude less than the accuracy of the 1977 experiment. Eisele also noted that near radio-pulsars, magnetic fields plausibly exist which could lead to an acceleration-induced correction to muon decay of almost 1%. But he correctly concluded that

... the most interesting part of this calculation surely consists not in any possible application like this but rather in the possibility in principle to verify the clock hypothesis in this special case [the $g = 2$ CERN experiment] with the help of an accepted physical theory. (Eisele 1987)

Something important is going on here. Unlike time dilation induced by uniform motion, which normally is understood to be independent of the constitution of the clock, the effects of acceleration will depend on the magnitude of the acceleration and the constitution of the clock. Fe^{57} nuclei and muons are much less sensitive to accelerations than “mechanical” clocks (like pendulum clocks), and the calculations of Sherwin and Eisele tell us why. They tell us why these microscopic clocks are capable of acting as odometers, or “waywisers”⁵ of time-like curves in relativistic space-time, even non-geodesic curves involving 3-accelerations of at least $10^{16}g$. But they also remind us that the general validity of the clock hypothesis for such accelerations is not a forgone conclusion. (For an interesting example of a quantum clock failing to satisfy the clock hypothesis, see Knox (2008).)

Before further discussion of the role of clocks in general relativity, it may be useful to remind ourselves of the origins of length contraction and time dilation in special relativity, without losing sight of the fact that special relativistic effects are ultimately grounded in general relativity.

2.3 Special Relativity

2.3.1 *Einstein’s Second Thoughts*

In special relativity it is clearly part of the theory that *uniformly moving* rods and clocks contract and dilate respectively in a fashion that does not depend on their constitution. Although it is a very remarkable part of the theory, this phenomenon of universality tends to be taken for granted today. It was far less obvious when in the aftermath of the 1887 Michelson-Morley experiment, variants of the experiment were performed using different materials for the rigid structure supporting the optical equipment (with the result of course that no difference in the null outcome was observed). In Einstein’s 1905 paper, this universality is simply assumed—it is built into the very operational significance he gave to the inertial coordinates and in particular their transformation properties. In addressing the question as to whether an explanation is actually available for the phenomenon, it is worth recalling the changing nature of Einstein’s own explanation of the “kinematical” phenomena of length contraction and time dilation by way of the Lorentz transformations.

There is an oddity in Einstein’s 1949 *Autobiographical Notes* (Einstein 1969) concerning his reconstruction of the development of special relativity. Einstein stressed, as he did throughout his life, the role that thermodynamics played as a

⁵See (Brown 2005, pp. 8 and 95).

methodological template in his 1905 thinking leading to *On the electrodynamics of moving bodies*. For reasons that have been much discussed in the literature, Einstein chose to base his new theory of space and time on principles which expressed “generalizations drawn from a large amount of empirical data that they summarize and generalize without purporting to explain” (Stachel 1998, p. 19)—in other words, principles like the phenomenological laws of thermodynamics. One of these principles was of course the principle of relativity, which in Einstein’s hands was restored to the *universal* version originally defended by Galileo and Newton. (The nineteenth century had seen the rise of ether theories of light and later of electromagnetism more generally, which raised widespread doubts as to whether the classical relativity principle should strictly apply to the laws of electrodynamics. Einstein’s aim was to banish such doubts, as Poincaré had actually urged before him, and even Lorentz and Larmor for up to second order effects.) Now it is important to emphasize that the formulation of this principle in 1905 does not presuppose the form of the (linear) coordinate transformations between inertial frames. After all, any hint that the Lorentz transformations are bound up in the expression of any one of Einstein’s postulates would have opened up the derivation in the kinematical part of the 1905 paper to the charge of circularity.

But note what Einstein says in his 1949 recollections. In relation to the general problem of describing moving bodies in electrodynamics, he wrote:

By and by I despaired of the possibility of discovering the true laws by means of constructive efforts based on known facts. The longer and the more despairingly I tried, the more I came to the conviction that only the discovery of a universal formal principle could lead us to assured results. The example I saw before me was thermodynamics. The general principle was there given in the theorem:⁶ the laws of nature are such that it is impossible to construct a *perpetuum mobile* (of the first and second kind). How, then, could such a universal principle be found? (Einstein 1969, p. 53)

The answer appears to be given, if at all, some paragraphs later:

The universal principle of the special theory of relativity is contained in the postulate: The laws of physics are invariant with respect to the Lorentz-transformations This is a restricting principle for natural laws, comparable to the restricting principle of the non-existence of the *perpetuum mobile* which underlies thermodynamics. (Einstein 1969, p. 57)

What is striking here is that Einstein is seemingly conflating the main consequence of his 1905 postulates (the principle of Lorentz covariance of the fundamental laws of physics) with one of the postulates themselves, namely the phenomenological relativity principle, the principle so beautifully brought out in Galileo’s famous thought experiment of the moving ship in his *Dialogue concerning two chief world systems*, and expressed more succinctly in Corollary V of the Laws in Newton’s *Principia*.⁷

⁶It is noted in (Brown 2005, p. 71, footnote 8) that the word “theorem” might be more happily translated from the original German as “statement”.

⁷In his 1949 discussion, Einstein clearly appreciates the difference between the two principles; see (Einstein 1969, p. 57).

The principle of Lorentz covariance has a more awkward affinity with the laws of thermodynamics than the relativity principle. Although it (Lorentz covariance) was widely accepted as a key element within the electromagnetic world picture by 1905, its plausibility for most theorists rested on the empirical success of a “constructive” theory whose strict validity Einstein himself doubted as a result of his light quantum hypothesis. More importantly, Lorentz covariance is far from the kind of phenomenological principles that characterise (pre-Carathéodory) versions of thermodynamics. The prohibition of perpetual motion machines of various kinds seems a higher-order constraint than the principle that the equations governing the mechanics underlying such machines—of which thermodynamics purports to be silent—satisfy a certain kind of boost symmetry.

Why this lapse on Einstein’s part? I wonder if it was not because of the misgivings he had about the way he formulated his 1905 paper, misgivings which grew throughout his life. First, there is little doubt that right from the beginning he was aware of the limited explanatory power of what he called “principle theories” like thermodynamics. Secondly, when he confessed in 1949 to having committed in 1905 the “sin” of treating rods and clocks as primitive entities, and not as “moving atomic configurations” subject to dynamical analysis, he was merely repeating a point of self-correction he made in 1921. Finally, it is fairly clear that Einstein was increasingly unhappy with the central role that electrodynamics, and in particular the behaviour of light, played in his 1905 paper.

This last aspect of Einstein’s reasoning brings us to the main point of this subsection. Einstein wrote in 1935:

The special theory of relativity grew out of the Maxwell electromagnetic equations. ... [but] the Lorentz transformation, the real basis of special-relativity theory, in itself has nothing to do with Maxwell theory. (Einstein 1935).

Similarly, in a 1955 letter to Born, Einstein would write that the “Lorentz transformation transcended its connection with Maxwell’s equations and has to do with the nature of space and time in general”. He went on to stress that “the Lorentz-invariance is a general condition for any physical theory.” (Born et al. 1971, p. 248). What is clear is that for the mature Einstein, the principle of Lorentz covariance, which applies to all the non-gravitational interactions, not just electrodynamics, is the heart of special relativity.⁸ In stressing this point, Einstein was distancing himself from his formulation of 1905 with its emphasis on

⁸In 1940, Einstein wrote: “The content of the restricted relativity theory can accordingly be summarised in one sentence: all natural laws must be so conditioned that they are covariant with respect to Lorentz transformations.” (Einstein 1954, p. 329). It is worth recalling in this context the way Einstein described in his *Autobiographical Notes* the main contribution Minkowski made to relativity theory. It was not so much Minkowski’s ontological fusion of space and time into a single four-dimensional entity that Einstein praised, but his provision of a tensor calculus in which equations for the non-gravitational interactions are manifestly Lorentz covariant. For Einstein, Minkowski had done for relativity what Heaviside and others did for Maxwell theory when they introduced the three-vector formulation of electrodynamics (so that the physics is manifestly Euclidean covariant). Minkowski “showed that the Lorentz transformation ... is nothing but a

fundamental phenomenological postulates (one of which being the “constancy” of the speed of light relative to the “rest” frame). The principle of Lorentz covariance is still a restriction on fundamental laws, but it is not quite like any of the laws of thermodynamics. By putting emphasis on its primacy, Einstein is effectively saying that the phenomenological relativity principle is a consequence of something deeper.

2.3.2 *The Sense in Which Rods and Clocks Don’t “Measure”*

This last point was brought out nicely in the 1976 pedagogical work of John S. Bell (1976) in special relativity, which consisted essentially of an extension of Oliver Heaviside’s 1889 result that the electrostatic field generated by a distribution of charge undergoes a distortion, according to Maxwell’s equations, when the charge is put into motion. Heaviside’s result was the inspiration for G. F. FitzGerald’s speculation concerning the deformation of rigid bodies moving through the luminiferous ether, and Bell himself took inspiration from both Heaviside and the work of Joseph Larmor to calculate the effect of (slow) accelerative motion on a 2-dimensional atom consisting of a heavy positively charged particle being orbited by a negatively charged particle and modeled using Maxwell’s equations, the Lorentz force law and the relativistic formula linking the moving particle’s momentum with its velocity. Bell showed that the atom spatially contracts and its period dilates when it achieves a uniform speed, in accordance with relativistic predictions.⁹ But more importantly for our purposes, he demonstrated that there is a new system of variables associated with the moving atom in relation to which the atom is described in

rotation of the coordinate system in the four-dimensional space” (Einstein 1969, p. 59), an insight which in fact Poincaré had anticipated.

⁹Suppose one considers the possibility of modeling a rigid rod by way of an infinite crystal composed of ions held together by electrostatic forces, rather in the spirit of Lorentz’s 1892 model of a system of charges held together in unstable equilibrium. Then the dynamical analysis seems to lead (as it did in the Lorentz case; see Brown (2001)) to a certain motion-induced deformation, rather than a strict longitudinal contraction: the conformal covariance of the equations has not been broken. This point was brought home to me in discussions some years ago with Adrian Sutton and his then 4th year undergraduate project students in the Department of Physics at Imperial College London: H. Anwar, V. Venkataraman, A. Wiener, C. Chan, B. Lok, C. Lin, and G. Abdul-Jabbar. This group has been studying a constructive approach to length contraction, similar to that of Bell (1976), but in which the effects of motion are calculated in a classical model of the attractive interatomic forces in an infinite ionic crystal. The question that has been thrown up, as I see it, is whether in such models it is possible to obtain strictly longitudinal length contraction without introducing quantum mechanics. (In the Bell atomic model, the conformal symmetry is broken by appeal to the Lorentz force law for the orbiting charge.) This question also applies to another electrostatic model of a rigid rod provided by Miller (2010), which was brought to my attention after discussions with the Imperial College group. Here, the way the author effectively breaks the conformal symmetry in the electrodynamics is not entirely consistent with the constructive nature of his approach.

exactly the same way that the stationary atom was described relative to the original variables—and that the new coordinates were related to the original ones by a Lorentz transformation. Bell had effectively derived the relativity principle—or that application of it to the simple electrodynamic system under consideration—from the dynamics postulated to hold in the original frame.

Bell of course realized that a satisfactory version of his atomic model would need to be reformulated in quantum theoretical terms, and that at root all the work in the argument is being done by the principle of Lorentz covariance of the fundamental equations. But what is important is that Bell's "Lorentzian pedagogy" is entirely free of the sin of treating rods and clocks as primitive entities, and it does not regard the relativity principle as fundamental. And in using this pedagogy as a warning against "premature philosophizing about space and time", Bell was reminding us that the reason rods and clocks do what they do is not because of what they are moving through but because of the dynamical principles of their very constitution. (Ironically, this lesson was in part lost to the ether theorists like Heaviside, FitzGerald and Larmor who were so influential on him.)

Bell also realized that the principle of Lorentz covariance needs to be assumed for *all* the interactions governing the constitution of matter, not just the electromagnetic forces. Although he did not stress it, it is therefore a consequence of the approach he is advocating (as it was in the work of Swann (1941), who had already applied the principle in the context of quantum theory) that rods and clocks contract and dilate respectively independently of their constitution. If Einstein had started with the universal principle of Lorentz covariance, he would not have needed to assume the universal nature of kinematics—that the inertial coordinate transformations codify the behaviour of rods and clocks whatever they are made of.

This point leads us to consider the question as to what it means to say such entities "measure". Much work has been expended in quantum theory in order to specify what is meant by a measurement process, inspired in good part by the so-called "measurement problem"—the problem of accounting, solely within quantum theory itself, for the emergence of well-defined results in standard measurement procedures. Models of such procedures are given by specifying a certain interaction Hamiltonian governing the coupling between the microscopic "object" system and the (usually macroscopic) measurement device. As a result of this coupling, the final state of the apparatus becomes correlated, in the appropriate sense, with the initial state of the object system, the combined system being subject to the time-dependent Schrödinger equation. Nothing like *this* notion—correlation through coupling—is taking place when rods and clocks do what they do in relativity theory, either at rest or in motion.

Bell's 1976 message was a (probably unwitting) reiteration of the message from a number of earlier commentators. Even before Swann, Pauli, for example, had already stressed in 1921, in his magisterial survey of relativity theory (Pauli 1921), that moving rods and clocks would not contract and dilate in conformity with the special relativistic predictions, were it not for the fact that all the non-gravitational interactions which account for the cohesive forces between the micro-constituents of these bodies are governed by Lorentz covariant equations. These objects are not

measuring anything in the above strong sense; they are merely acting in accordance with the dynamical laws governing their internal make-up. They are not the analogue of thermometers of the space-time metric. They measure in the weaker sense that their behaviour *correlates* with aspects of space-time structure and that is why we are interested in them and use our best theories to build them¹⁰—but it is not because space-time *acts* on them in the way a heat bath acts on a thermometer, or the way a quantum system acts on a measuring device.

2.4 General Relativity

This last claim might be regarded as much more contentious in general relativity (GR) than in special relativity. After all, in special relativity the Minkowski metric is absolute and non-dynamical.¹¹ But in GR, the action-reaction principle seems gloriously restored, its reincarnation now involving matter and metric. But how does general relativity purport to predict that an ideal clock, for example, will act as a way-wiser of space-time without treating it as a primitive entity?¹²

It may be useful in answering this question to consider momentarily the case of alternative theories of gravity which feature more than one metric field, and in particular *bimetric* theories. For example, Rosen (1980), in an attempt to avoid

¹⁰Note that the distinction here between strong and weak measurement is not relevant to the issue of the accuracy of the measurement.

¹¹A possible objection to this reasoning might go as follows. The generally covariant formulation of any specially relativistic dynamics (such as Maxwellian electrodynamics), or more generally any dynamical theory within an absolute space-time background, flat or curved, leads to equations of motion in which the absolute structure appears in the equations. Such structure appears to be causally relevant; indeed a violation of the action-reaction principle seems to obtain. Space-time structure acts on matter, but not the other way round. However, demanding general covariance in the context of special relativity is like demanding that (“pure gauge”) electromagnetic vector and scalar potentials appear in the Schrödinger equation for a free particle. Just as it would be odd to say that such potentials are physically acting on the particle, arguably the “action” of space-time structure in special relativity is merely an artifact of the generally covariant formulation, which is ill-suited to the theory. (For further discussion of the purported violation of the action-reaction principle in special relativity theory, see Brown and Pooley (2004) and Brown (2005, section 8.3.1). For a treatment of Einstein’s appeal to the principle in extolling the virtues of general relativity, see Brown and Lehmkuhl (2016).

¹²It is curious how infrequently this issue is raised. A rare case was Dieks, writing in 1987: “... it should be emphasized that the general theory of relativity is a fundamental physical theory ... it can safely be said that constructs like macroscopic measuring rods and clocks cannot figure as essential elements in such a fundamental theory. ... [T]he behaviour of macroscopic bodies like rods and clocks should be explained on the basis of their microscopic constitution” (Dieks 1987, p. 15); see also (Dieks 1984.) However, the nature of this explanation as suggested by Dieks differs from what follows. I note that Fletcher (2013) provides a theorem showing that, for any timelike curve in any spacetime, there is a light clock that measures the length of the curve as given by the metric to arbitrary accuracy. The proof of course assumes that light propagates along null geodesics, which is a consequence of the Einstein equivalence principle (see below).

the singularities that appear in standard GR, introduced besides the $g_{\mu\nu}$ field that describes gravity, a non-dynamical metric field $\gamma_{\mu\nu}$ of constant curvature which serves to define a fundamental rest frame of the universe. More recently, Bekenstein (2004) developed a bimetric theory which is a covariant version of Milgrom's MOND program, designed principally to account for the anomalous rotation curves in galaxies without an appeal to dark matter. Bekenstein's Tensor-Vector-Scalar theory (TeVeS) incorporates two metric fields.¹³ The first, represented by $g_{\mu\nu}$, has as its free Lagrangian density the usual Hilbert-Einstein Lagrangian. The second, represented by $\tilde{g}_{\mu\nu}$, can be expressed as a deformation—a disformal transformation—of $g_{\mu\nu}$ according to a formula that depends on fundamental vector and scalar fields postulated to exist alongside $g_{\mu\nu}$. What is of interest to us in the present context is not whether these bimetric theories are true, but how it is in each theory that the usual rods and clocks end up surveying (at most) only one of the two metrics, as they must.

Take TeVeS. It is critical to the enterprise that the metric which is assigned chronometric significance is the “less basic” $\tilde{g}_{\mu\nu}$, not $g_{\mu\nu}$. It is $\tilde{g}_{\mu\nu}$ that is “measured” by rods and clocks, and whose conformal structure is traced out by light rays and whose time-like geodesics are the possible worldlines of free bodies. Bekenstein is clear as to how this is realized: $\tilde{g}_{\mu\nu}$ is “delineated by matter dynamics” in the right way. The issue has to do with the way the usual matter fields are postulated to couple to $\tilde{g}_{\mu\nu}$ (and therefore to $g_{\mu\nu}$). In short, the *Einstein equivalence principle* is defined relative to $\tilde{g}_{\mu\nu}$ in the theory. Bekenstein, in establishing the operational significance of $\tilde{g}_{\mu\nu}$ in TeVeS, is simply doing what is done—though often with less emphasis—in relation to $g_{\mu\nu}$ in standard GR.

Let us consider an arbitrary event P in space-time. There exist in the neighbourhood of P locally inertial (Lorentz) coordinates, such that at P the first derivatives of $g_{\mu\nu}$ vanish and $g_{\mu\nu} = \eta_{\mu\nu}$ (read $\tilde{g}_{\mu\nu}$ for $g_{\mu\nu}$ in TeVeS). The Einstein equivalence principle has two components. The first (universality) states that the fundamental laws for *all* the non-gravitational interactions involving matter fields take their simplest form at P relative to these local coordinates. The second (minimal coupling) states that this form is the special relativistic form. As Ohanian put the principle: “At each point of space-time it is possible to find a coordinate transformation such that the gravitational field variables can be eliminated from the field equations of matter” (Ohanian 1977). It ensures, in the words of Misner, Thorne and Wheeler, that “. . . in any and every local Lorentz frame, anywhere and anytime in the universe, all the (nongravitational) laws of physics must take on their familiar special relativistic forms” (Misner et al. 1973, p. 386). Such, at any rate, are typical formulations of the principle; however, it should not be overlooked—as

¹³For a recent review of TeVeS, see Skordis (2009). It has been argued (Zlosnik et al. 2006) that TeVeS is not a true bimetric theory. First, it can be shown to be equivalent to a (mathematically more complicated) Tensor-Vector theory involving just the single metric $\tilde{g}_{\mu\nu}$ in the total action. More significantly, these authors claim that tensor gravity waves propagate along the same light cone as electromagnetic ones. But this claim conflicts with the analysis of TeVeS and its generalizations by Skordis (2006, 2008, 2009).

indeed the latter authors stress—that curvature-related terms will generally appear in any second-order equations involving the non-gravitational interactions. (For more details, see Brown and Read (2016)).

It follows nonetheless from the Einstein equivalence principle that the fundamental equations governing all the non-gravitational interactions are locally Lorentz covariant,¹⁴ and hence that in so far as we can ignore the effects of tidal forces on rods and clocks, they will behave in conformity with the predictions of special relativity, as stressed in Section 2.3. (In the case of accelerating rods and clocks, the kinds of qualifications raised in Section 2.2 will of course still be relevant to the issue of the universality of their behaviour). And so the phenomenological relativity principle (or rather its local variant) for the non-gravitational interactions is itself a consequence of the Einstein equivalence principle.

In stressing the role of this principle in accounting for the behaviour of rods and clocks in general relativity, an observation due to James L. Anderson in his remarkable 1976 book *Principles of Relativity Physics* comes to mind. Anderson claims that of the two components above of the principle, the first (which implies that measurements on any physical system will determine the same affine connection) is essential to GR—to any metric theory of gravity—and the second is not (Anderson 1967, section 10-2). Be that as it may, it is worth recalling that the Einstein equivalence principle is not a strict consequence of the form of Einstein’s field equations, any more than the signature of the metric is, or any more than the Galilean covariance of Newtonian mechanics follows from the strict form of Newton’s laws of motion. (The further assumptions that all Newtonian forces are velocity-independent like the gravitational force, and similarly that inertial masses are velocity-independent, are jointly required for the Galilean covariance of Newton’s laws). If the metric field is to be considered a property of space-time, in the sense of Will (see above), it requires a very non-trivial dynamical assumption to be made over and above postulating the field equations, which is roughly that special relativity, properly understood, holds locally.

But just because the metric field *can* be so considered, it is not clear it *should* be. Or rather, it is not clear that the geometrical interpretation of $g_{\mu\nu}$ is intrinsic to its dynamical role in GR, particularly when one considers a non-trivial space-time completely free of matter and hence rods and clocks. Even in the general case, the notion that the $g_{\mu\nu}$ field is the fabric of space-time, rather than a field in space-time, may be misleading.¹⁵ It may serve to hinder recognition of the possibility that Einstein gravity is an emergent phenomenon, for example in the sense that it is a

¹⁴A gravitational theory that violates local Lorentz covariance is due to Jacobson et al. (2001). It contains a time-like unit vector field which serves to pick out a preferred frame. I take it the appearance of this field in the equations governing the matter fields is ruled out by the second component of the Einstein equivalence principle.

¹⁵A different analysis leading to the same conclusion was provided by James L. Anderson (1999). Anderson also argued that the (approximately) metrically-related behaviour of clocks can be derived from the dynamical assumptions of GR, in the same way that the motion of a free test particle can be derived. He claimed that this becomes particularly clear in the approximation

consequence of the specific dynamics of an evolving fundamental 3-geometry,¹⁶ or in the stronger sense that the field equations are analogous to the equations of fluid dynamics, which emerges from molecular physics as a low-momentum long-distance approximation.¹⁷ These non-standard approaches may not prove to be correct, but they—and particularly the latter view which goes some way to deriving the Lorentzian signature of the metric, and furthermore calls into question the program of quantizing gravity—should not be dismissed lightly. At the very least, the view that $g_{\mu\nu}$ is space-time structure may serve to hide from sight the fact that the Einstein equivalence principle is a highly non-trivial part of GR, and that the universality of the principle of Lorentz covariance incorporated therein is arguably mysterious—by which I mean explanation-seeking—particularly in the absence of a strict theoretical unification of all the non-gravitational interactions.¹⁸

It is noteworthy that for whatever reasons, Einstein at the end of his life sounded clear notes of caution on the question of the interpretation of the metric field. In 1948, he wrote in a letter to Lincoln Barnett:

I do not agree with the idea that the general theory of relativity is geometrizing Physics or the gravitational field. The concepts of Physics have always been geometrical concepts and I cannot see why the g_{ik} field should be called more geometrical than f.[or] i.[nstance] the

scheme developed to address the problem of motion in GR due to Einstein, Infeld and Hoffmann in 1939 and 1940. I repeat Anderson's concluding remarks:

In this paper I have argued that a metric interpretation is not needed in general relativity and that the purposes for which it was originally introduced, i.e., temporal and spatial measurements and the determination of geodesic paths, can be all be derived from the field equations of this theory by means of the EIH [Einstein, Infeld and Hoffman] approximation scheme. As a consequence, the only *ab initio* space-time concept that is required is that of the blank space-time manifold. In this view what general relativity really succeeded in doing was to eliminate geometry from physics. The gravitational field is, again in this view, just another field on the space-time manifold. It is however a very special field since it is needed in order to formulate the field equations for, what other fields are present and hence couples universally with all other fields.

¹⁶See the reconstruction of GR due to Barbour et al. (2002) based on a dynamical 3-geometry approach and inspired by Mach's relational reasoning.

¹⁷The emergent approach discussed by Barceló et al. (2001) relies on classical considerations related to the so-called "analog models" of GR to motivate the existence of the Lorentzian metric field, and effective theories arising out of the one-loop approximation to quantum field theory to generate the dynamics (and in particular the familiar Hilbert-Einstein term in the effective action) in the spirit of Sakharov's 1968 notion of induced gravity. Such an approach clearly calls into question the appropriateness of quantizing gravity. For more recent developments along similar lines, but now based on a potentially deep connection between the field equations for the metric and the thermodynamics of horizons, see Padmanabhan (2007, 2008).

¹⁸The above-mentioned work of Barbour et al. (2002) was originally thought to provide a derivation of the Einstein equivalence principle. However, careful further analysis by Edward Anderson has cast doubt on this claim; for details see Anderson (2007). It should be mentioned that the validity of the Einstein equivalence principle seems to be genuinely mysterious in the case of the case of the emergent gravity approach; see Barceló et al. (2001, section 4).

electro-magnetic field or the distance of bodies in Newtonian Mechanics. The notion comes probably from the fact that the mathematical origin of the g_{ik} field is the Gauss-Riemann theory of the metrical continuum which we are wont to look at as a part of geometry. I am convinced, however, that the distinction between geometrical and other kinds of fields is not logically founded.¹⁹

And in Einstein's 1949 *Autobiographical Notes*, when attempting to justify his "sin" of treating rods and clocks as primitive in his 1905 relativity paper by appealing to the (then) lack of understanding of the microphysics of matter, he stressed:

But one must not legalize the mentioned sin so far as to imagine that intervals are physical entities of a special type, intrinsically different from other physical variables ("reducing physics to geometry", etc.). (Einstein 1969, p. 61)

Acknowledgements I wish to thank David Rowe, Tilman Sauer and Scott Walter for the kind invitation to contribute to this volume. I am grateful to Dennis Lehmkuhl, David Rowe and Scott Walter for comments that led to improvements in the paper. I thank Norbert Straumann for bringing to my attention the 1987 work by Anton Eisele, and for further correspondence. I also benefited from discussions with Eleanor Knox, Constantinos Skordis and George Svetlichny, as well as Adrian Sutton and his project students (see note 9).

References

- Anderson, E. (2007). On the recovery of geometrodynamics from two different sets of first principles. *Studies in History and Philosophy of Modern Physics*, 38, 15–57.
- Anderson, J. L. (1967). *Principles of relativity physics*. New York: Academic Press Inc.
- Anderson, J. L. (1999). Does general relativity require a metric. arXiv:gr-qc/9912051v1.
- Bailey, J., Borer, K., Combley, F., Drumm, H., Farley, F. J. M., Field, J. H., et al. (1977). Measurements of relativistic time dilation for positive and negative muons in a circular orbit. *Nature*, 268, 301–305.
- Barbour, J., Foster, B. Z., & Murchadha, N. Ó. (2002). Relativity without relativity. *Classical and Quantum Gravity*, 19(12), 3217–3248.
- Barceló, C., Visser, M., & Liberati, S. (2001). Einstein gravity as an emergent phenomenon? arXiv:gr-qc/0106002v1.
- Bekenstein, J. D. (2004). Relativistic gravitation theory for the modified Newtonian dynamics paradigm. *Physical Review D*, 70, 083509.
- Bell, J. S. (1976). How to teach special relativity. *Progress in Scientific Culture*, 1, reprinted in J.S. Bell, *Speakable and Unsayable in Quantum Mechanics*, Cambridge: Cambridge University Press (2008), pp. 67–80.
- Born, M., Born, H., & Einstein, A. (1971). *The Born-Einstein letters*. London: Macmillan.
- Brown, H. R. (2001). The origins of length contraction: I. The FitzGerald-Lorentz deformation hypothesis. *American Journal of Physics*, 69, 1044–1054. arXiv:gr-qc/0104032; PITT-PHIL-SCI 218.

¹⁹I am grateful to Dennis Lehmkuhl for recently bringing this letter to my attention. It is reproduced in the preface by John Stachel in (Earman et al. 1977, p. ix). An enlightening account of Einstein's misgivings about the geometric nature of general relativity is given by Lehmkuhl (2014). Finally, a fuller treatment of most of the main arguments made in sections 3 and 4 of the present paper is given by Brown (2005).

- Brown, H. R. (2005). *Physical relativity. Space-time structure from a dynamical perspective*. Oxford: Oxford University Press. The 2007 paperback edition removes a number of errata from the original.
- Brown, H. R., & Lehmkuhl, D. (2016). Einstein, the reality of space, and the action-reaction principle. In P. Ghose (Ed.), *Einstein, Tagore and the nature of reality* (pp. 9–36). London: Routledge. Preprint <http://philsciarchive.pitt.edu/id/eprint/9792>.
- Brown, H. R., & Pooley, O. (2004). Minkowski space-time: A glorious non-entity. arXiv:physics/0403088; PTT-PHIL-SCI 1661. A revised version appeared in *The Ontology of Space, I*, D. Dieks (Ed.), Amsterdam: Elsevier, 2006, pp. 67–89.
- Brown, H. R., & Read, J. (2016). Clarifying possible misconceptions in the foundations of general relativity. *American Journal of Physics*, 84(5), 327–334.
- Carroll, S. (2004). *Spacetime and geometry. An introduction to general relativity*. San Francisco: Addison-Wesley.
- Cranshaw, T. E., Schiffer, J. P., & Whitehead, A. B. (1960). Measurement of the gravitational red shift using the Mössbauer effect in Fe⁵⁷. *Physical Review Letters*, 4, 163–164.
- Dieks, D. (1984). On the reality of the Lorentz contraction. *Zeitschrift für allgemeine Wissenschaftstheorie*, 15, 330–42.
- Dieks, D. (1987). Gravitation as a universal force. *Synthese*, 73, 381–39.
- Earman, J., Glymour, C., & Stachel, J. (Eds.). (1977). *Minnesota studies in the philosophy of science, volume VIII: Foundations of space-time theories (proceedings)*. Minnesota: University of Minnesota Press.
- Eddington, A. S. (1966). *Space, time and gravitation. An outline of the general theory of relativity*. Cambridge: Cambridge University Press.
- Einstein, A. (1935). Elementary derivation of the equivalence of mass and energy. *Bulletin of the American Mathematical Society*, 41, 223–30.
- Einstein, A. (1954). The fundamentals of theoretical physics. In *Ideas and opinions* (pp. 323–335). New York: Bonanza Books.
- Einstein, A. (1969). Autobiographical notes. In P. A. Schilpp (Ed.), *Albert Einstein: Philosopher-scientist* (vol. 1, pp. 1–94). Chicago: Open Court.
- Eisele, A. (1987). On the behaviour of an accelerated clock. *Helvetica Physica Acta*, 60, 1024–1037.
- Fletcher, S. C. (2013). Light clocks and the clock hypothesis. *Foundations of Physics*, 43(11), 1369–1383.
- Hay, J. J., Schiffer, J. P., Cranshaw, T. E., & Egelstaff, P. A. (1960). Measurement of the red shift in an accelerated system using the Mössbauer effect in Fe⁵⁷. *Physical Review Letters*, 4, 165–166.
- Hentschel, K. (1996). Measurements of gravitational redshift between 1959 to 1971. *Annals of Science*, 53, 269–295.
- Jacobson, T., & Mattingly, J. (2001). Gravity with a dynamical preferred frame. *Physical Review D*, 64, 024028.
- Josephson, B. D. (1960). Temperature-dependent shift of γ rays emitted by a solid. *Physical Review Letters*, 4, 341–342.
- Knox, E. (2008). Flavour-oscillation clocks and the geometricity of general relativity. arXiv:gr-qc/0809.0274v1.
- Lehmkuhl, D. (2014). Why Einstein did not believe that general relativity geometrizes gravity. *Studies in History and Philosophy of Modern Physics*, 46, 316–326.
- Miller, D. J. (2010). A constructive approach to the special theory of relativity. *American Journal of Physics*, 78, 633–638.
- Misner, C. W., Thorne, K. S., & Wheeler, J. A. (1973). *Gravitation*. San Francisco: Freeman & Co.
- Ohanian, H. C. (1977). What is the principle of equivalence? *American Journal of Physics*, 45, 903–909.
- Padmanabhan, T. (2007). Gravity as an emergent phenomenon: A conceptual description. arXiv:gr-qc/0706.1654v1.
- Padmanabhan, T. (2008). Gravity: The inside story. *General Relativity and Gravitation*, 40, 2031–2036.

- Pauli, W. (1921). Relativitätstheorie. In A. Sommerfeld (Ed.), *Encyklopädie der mathematischen Wissenschaften, mit Einschluss ihrer Anwendungen, Vol. 5: Physik*. Leipzig: Teubner. English translation: *Theory of Relativity*, New York: Dover, 1981.
- Pound, R. V., & Rebka, G. A., Jr. (1960a). Apparent weight of photons. *Physical Review Letters*, 4, 337–341.
- Pound, R. V., & Rebka, G. A., Jr. (1960b). Variation with temperature of the energy of recoil-free gamma rays from solids. *Physical Review Letters*, 4, 274–275.
- Rosen, N. (1980). General relativity with a background metric. *Foundations of Physics*, 10, 637–704.
- Schild, A. (1960). The equivalence principle and red-shift measurements. *American Journal of Physics*, 28, 778–780.
- Sherwin, C. W. (1960). Some recent experimental tests of the ‘clock paradox’. *Physical Review*, 120, 17–24.
- Skordis, C. (2006). Tensor-vector-scalar cosmology: Covariant formalism for the background evolution and linear perturbation theory. *Physical Review D*, 74, 103513.
- Skordis, C. (2008). Generalizing tensor-vector-scalar cosmology. *Physical Review D*, 77, 123502.
- Skordis, C. (2009) The tensor-vector-scalar theory and its cosmology. arxiv.org/abs/0903.3602v1.
- Stachel, J. (1998). Introduction. In J. Stachel (Ed.), *Einstein’s miraculous year. Five papers that changed the face of physics* (pp. 3–27). Princeton: Princeton University Press.
- Swann, W. F. G. (1941). Relativity, the Fitzgerald-Lorentz contraction, and quantum theory. *Reviews of Modern Physics*, 13, 190–196.
- Wilkie, T. (1977). The twin paradox revisited. *Nature*, 268, 295–296.
- Will, C. M. (2001). The confrontation between general relativity and experiment. <http://www.livingreviews.org/lrr-2001-4>.
- Zlosnik, T. G., Ferreira, P. G., & Starkmann, G. D. (2006). The vector-tensor nature of Bekenstein’s relativistic theory of modified gravity. *Physical Review D*, 74, 044037. arXiv:gr-qc/0606039v1.

Chapter 3

Hilbert on General Covariance and Causality



Katherine Brading and Thomas Ryckman

3.1 Introduction

Hilbert's work on generally covariant physics in 1915 led him to diagnose a tension between general covariance and causality, and to seek its resolution. In an earlier paper in this series (Brading and Ryckman 2009), we presented Hilbert's reconsideration of the status of causality in the light of general covariance as it unfolds in Hilbert's First and Second Communications on the Foundations of Physics (Hilbert 1915a,b, 1917). In our paper, we claim that Hilbert's "causality problem" and the resolution he offers differ from the (in)famous "hole argument" and its resolution, due to Einstein (see also Brading and Ryckman (2008, section 7)). The questions and feedback that we continue to receive when discussing this research have made it clear that a supplementary note would be valuable, giving further details of the differences between Hilbert's "causality problem" and that of Einstein, and also making explicit the relationship between Hilbert's proposed resolution and how we think about general covariance and causality in General Relativity today. The purpose of this paper is to address these two points.

We begin with a review of Einstein's "causality problem" and the solutions that he offers (Sections (3.2) and (3.3)). We then discuss the evolution of Hilbert's "causality problem" through the First and Second Communications (Sections (3.4) and (3.5)), before addressing (in Section (3.6)) the resolution that he offers in the Second Communication, including its relationship to how we think about these things today. Hilbert's "causality problem" has both a mathematical

K. Brading (✉)

Department of Philosophy, Duke University, Durham, NC 27708, USA
e-mail: katherine.brading@duke.edu

T. Ryckman

Department of Philosophy, Stanford University, Stanford, CA 94305, USA
e-mail: tryckman@stanford.edu

and an epistemological face, and while the mathematical problem and its resolution are standard fare in General Relativity today, his epistemological discussions remain largely unknown. We end by comparing Einstein's "causality problem" with that of Hilbert, and here make the case that Hilbert was never a victim of Einstein's "hole argument" (see Section (3.7)).

3.2 Einstein's "Causality Problem"

When Einstein was lecturing in Göttingen during the summer of 1915, he still believed that generally covariant field equations were not to be had. He had two arguments for this, one is the so-called "hole argument" (see Stachel (1989)), and the other has to do with energy conservation.

Einstein's hole argument has been the subject of detailed consideration in the history and philosophy of general relativity literature.¹ In the argument, Einstein considers a region of spacetime in which there are no matter fields present (the "hole"). He then shows that, in a generally covariant theory, no amount of data about the values of the matter and gravitational fields (or the metric) *outside* the hole is sufficient such that, when combined with the field equations, the values of the gravitational field *inside* the hole are uniquely determined. This was unacceptable to Einstein: motivated by what he would later refer to as "Mach's principle", he was searching for a theory in which the matter fields plus the field equations would uniquely determine the metric.² Thus, the hole argument can be understood as posing a kind of "causality problem" for any generally covariant theory.

Einstein formulated the hole argument as a *post hoc* justification for his failure to find generally covariant field equations. His thinking about energy conservation had led him to conclude that we need to restrict the covariance class of our theory. The conclusion of the hole argument was that no generally covariant theory will be physically acceptable. Using energy-momentum conservation to arrive at four non-generally covariant conditions, Einstein restricted the covariance properties of his field equations and thereby restored "causality". This seemed satisfactory: Einstein had an argument for why no generally covariant theory could be physically possible (the hole argument), and he had conditions limiting the covariance class of his field equations that were motivated by physical considerations (energy conservation).³

While we do not know for sure what Einstein said in his 1915 Göttingen lectures, they were surely the catalyst for Hilbert's First and Second Communications

¹In addition to Stachel (1989), see also Norton (1984, pp. 286-91), Norton (1993, sections 1-3), (Ryckman 2005, section 2.2.2), and references therein.

²For more on Einstein's (mis)appropriation of Mach's principle, see Barbour (2005).

³(Einstein and Grossmann 1914). See Janssen and Renn (2007), for the details of the story.

(Hilbert 1915a,b, 1917). Hilbert is explicit in attributing the idea of generally covariant physics to Einstein, and it is reasonable to infer that two key features of Einstein's work on gravitational theory (concerning a conflict between general covariance and causality, and the use of conservation of energy to resolve the problem) were included in his lectures. It is also reasonable to infer that they were picked up by Hilbert, then to appear in the December Proofs of his First Communication (Hilbert 1915a).⁴ However, as we will see, Hilbert had a different understanding from Einstein of the problem of causality raised by general covariance, and also, therefore, used the considerations about energy conservation rather differently.

The appeal to energy conservation to address problems of causality was consigned to the scrap heap before the year was out. On the 2nd of December 1915, Einstein published the familiar Einstein Field Equations of General Relativity which are, of course, generally covariant; energy conservation no longer restricts the covariance properties of the field equations.

Einstein later "solved" the "problem of causality" posed in his hole argument via his "point coincidence argument". As we hope to make clear in what follows, it is this solution that helps to pinpoint Einstein's own "causality problem" and to make vivid the differences between this and Hilbert's "causality problem".

3.3 Einstein's "Point Coincidence Argument"

As is now well-documented, Einstein extricated himself from the conclusion of the hole argument by means of his so-called "point coincidence argument".⁵ However, this resolution of the difficulty was only obliquely expressed in print in the canonical presentation of the new theory published on 11 May 1916 (Einstein 1916). A passage in section 3 of that paper was first identified as the "point-coincidence argument" by Stachel (1989), and it presents a puzzle: it begins with a declaration that the requirement of general covariance removes "the last remnant of physical objectivity from space and time", but in support of this apparently ontological conclusion offers what seems to be a suspiciously epistemological argument. The first premise states that all of our spacetime observations and measurements ultimately amount to a determination of spacetime coincidences. As illustrative examples of this premise, Einstein cites the meeting of two or more material points, and even the coincidence between a pointer and the marks on a dial. The second premise concerns the role of coordinate systems, and is the suggestion that the introduction of a coordinate system merely facilitates the description of the totality

⁴The "December Proofs" (Hilbert 1915a) were recently discovered (see Corry, L., J. Renn, and J. Stachel (1997) and contain significant differences from the published version (Hilbert 1915b). For discussion of these differences, and differing opinions on their significance, see Renn and Stachel (1999), and Sauer (1999, 2005).

⁵See Stachel (1989), and also Norton (1993, section 3.5), and Ryckman (2005, p. 21).

of such coincidences. But, since two coincident point events (described by identical coordinates in a given coordinate system) will remain coincident in a new coordinate system (arrived at from the first by an arbitrary coordinate transformation), we have no reason to prefer one system of coordinates to any other. Thus, we arrive at the requirement of general covariance.

The verificationist flavor of this argument was widely hailed by Machians (such as Phillip Frank) and positivists of various stripes. Following Stachel (and in the light of much later Einstein texts), we suppose a more charitable gloss can be given to the point coincidence argument. What Einstein should have said is that the discordant conclusion stemming from the hole argument (that generally covariant field equations lead to indeterminism) no longer goes through once it has been recognized that systems of spacetime coordinates have no metrical or other physical significance, but serve as essentially arbitrary labels for spacetime points. Thus the supposedly distinct solutions generated by given matter sources can now be recognized as being merely different mathematical descriptions of the same physical state of affairs. It is this realization that enables Einstein to evade the causality problem posed by the hole argument.

3.4 Hilbert's "Causality Problem", 1915

Throughout the First Communication, both proofs and published version (Hilbert 1915a,b), and the Second Communication (Hilbert 1917) Hilbert never wavers from his commitment to general covariance. As we have argued (Brading and Ryckman 2008, 2009), Hilbert saw profound epistemological significance in general covariance, and sought to explore the consequences of adopting it as an axiom of fundamental physics.

Already in the proofs, Hilbert makes clear the implications of generally covariant physics for considerations of causality, as he understood them. He claims that *any* generally covariant theory will face a problem of mathematical underdetermination, stating explicitly in his Theorem 1 that for a system of n Euler-Lagrange differential equations in n variables obtained from a generally covariant action integral, there will be only $n - 4$ equations for the n variables. Hilbert then argues as follows:

Therefore, if we want to preserve the determinate character of the fundamental equations of physics according to Cauchy's theory of differential equations, the requirement of four additional non-invariant equations supplementing [the field equations] is essential. (Hilbert 1915a, p. 4)

In this extract we see two things clearly stated: the first is Hilbert's "causality problem", and the second is his proposal for its resolution. Hilbert explicitly states that the causality problem associated with general covariance concerns Cauchy determination. The question is whether a generally covariant theory admits of a well-posed Cauchy problem, and Theorem 1 suggests that it does not. In the context of spacetime theory and a system of second-order partial differential

equations on that spacetime, a well-posed Cauchy problem requires that the initial data assignments to the unknown field functions and their first (time) derivatives on a spacelike hypersurface determine the second time derivatives of the given field quantities, and thereby unique solutions off the hypersurface (for appropriate regions). Hilbert's Theorem 1 pinpoints failure of Cauchy determination as a consequence of general covariance. This is Hilbert's "problem of causality" in 1915.

Distinct from this diagnosis of a causality problem is the solution Hilbert offers in the proofs: the addition of four further equations. Hilbert followed Einstein in using energy conservation to provide these additional equations, but his purpose was somewhat different. For Hilbert, general covariance retains its axiomatic status, and the field equations remain generally covariant; but, for the sole purpose of meeting the mathematical requirement of Cauchy determination within this generally covariant structure, additional conditions (deriving from energy conservation) are imposed.⁶

With the publication of the generally covariant Einstein Field Equations (for which, of course, Einstein no longer uses energy conservation to restrict the covariance properties of the field equations) Hilbert had to abandon this solution to his causality problem. And, indeed, when we look at the published version of the First Communication, this whole application of energy conservation has gone. But Theorem 1 is still there in the published version: so the causality problem is still there, but now Hilbert has no solution for it.

3.5 Hilbert's "Causality Problem", 1917

Hilbert's Second Communication (Hilbert 1917) includes a new treatment of the causality problem originally posed in the First Communication, embedding and developing the original mathematical problem in an explicit epistemological context. As we discuss in what follows, Hilbert presents his causality problem as an apparent conflict between the *axiom* of general covariance and our *experience* of the world as causally ordered and causally determinate. This makes explicit the *epistemological* aspect of the problem, which in the First Communication had appeared under a predominantly *mathematical* guise. For Hilbert the deep problem is the epistemological problem.

It is clear that from the outset Hilbert saw deep epistemological significance in general covariance. He saw the adoption of general covariance as an important step towards removing the contributions of human subjectivity from the conceptual

⁶Some commentators have mistakenly asserted that Hilbert's equations are not generally covariant. On this issue we wholeheartedly support Ohanian's recent statement when he writes: "The fact is that Hilbert's variational equations are covariant, but he supplements them, correctly, by extra, noncovariant, coordinate conditions that are needed to make the solution unique, as is well known to anybody who has ever tried to construct a solution of the Einstein equations." (Ohanian 2008, p. 355 (n. 56 to p. 221)).

structure of physics; specifically, by making it independent of the way in which world-points are designated (through coordinates). Thus, immediately following the statement of his axiom of general covariance in his First Communication, he writes that this axiom is: “the simplest mathematical expression of the demand that the inter-linking of the potentials $g_{\mu\nu}, q_s$ is by itself entirely independent of the way one chooses to label the world’s points by means of world parameters.” This statement appears in both the proofs and the published version (Hilbert 1915a, p. 990); (Hilbert 1915b, p. 1004). The point is repeated again in Hilbert (1919/1992, p. 49), and in 1921 Hilbert describes the move to general covariance as an emancipation from “the *subjective* moments of human *intuition* with respect to space and time” and “a radical elimination of *anthropomorphic* slag” (cited in Majer 1995, p. 284). This understanding of general covariance leads to a corresponding epistemological aspect of the “causality problem”, because the requirement of causality appears to be inconsistent with the emancipation achieved by general covariance.

As in the First Communication, the tension between general covariance and causality is given a precise mathematical characterization: Hilbert points out that general covariance leads to a mathematical problem with respect to Cauchy determination. New to the mathematical discussion is Hilbert’s observation that arbitrary point transformations (diffeomorphisms) do not respect the relation of cause and effect among world points lying on the same timelike curve: they allow transformations that reverse the temporal order of “cause” and “effect” or place them in spacelike relation. Thus, the mathematical problem now concerns both causal ordering as well as Cauchy determination.

Hilbert’s position is that we need to reconcile the general covariance of the conceptual structure of physics with our experience of the world as causal (both causally ordered, and causally determinate in the sense of Cauchy determination). That is, we must be able to recover the world *as we subjectively experience it* from the generally covariant structure of *objective* physics. This is Hilbert’s “causality problem” in 1917.

3.6 Hilbert’s 1917 Resolution of His “Causality Problem”

To address his “causality problem”, Hilbert begins by introducing the notion of “proper coordinate systems”: by definition, transformations among such coordinate systems respect the distinction between spacelike and timelike coordinate axes, preserve the temporal ordering of cause and effect, and ensure Cauchy determination. If we restrict ourselves to the use of proper coordinate systems, we will extract causally determinate structures appropriate for expressing our *experience* of the world, from the generally covariant conceptual structure.

The mathematical aspect of Hilbert’s “causality problem”, of achieving Cauchy determination, is solved via appeal to “proper coordinate systems”. As Stachel (1992) states, Hilbert was the first to discuss the Cauchy problem for the Einstein equations. The solution Hilbert offers in his Second Communication (Hilbert 1917)

makes explicit use of Gaussian coordinates, which put the remaining 10 equations into Cauchy normal form. While Stachel also points out that many of the subtleties and difficulties posed by general covariance escaped Hilbert in 1917, the problem that Hilbert posed and the general method for its solution continue to be a part of standard practice in General Relativity.⁷

The epistemological face of Hilbert's "causality problem" is less familiar, and it is to this that Hilbert turns his attention having addressed the mathematical issue. We noted above that, for Hilbert, general covariance has profound epistemological significance as *the* criterion of physical objectivity, enabling us to eliminate the "anthropomorphic slag" associated with preferred choices of coordinate systems. What Hilbert is looking for is an account of the relationship between the physically objective world as expressed by generally covariant field equations, and our subjective experience of the world as causally ordered and causally determinate. According to our reconstruction (see Brading and Ryckman 2008, 2009), the approach he takes is to consider the status of the statements that we make reporting our experiences of the world as causally ordered and determinate. In the Second Communication Hilbert explicitly insists that physically meaningful propositions in physics *must* have a generally covariant formulation. However, this condition is not by itself sufficient for physical meaningfulness. In the light of "the principle of causality" (Hilbert 1917, p. 1024), a second condition must be added, according to which when such a proposition is expressed with respect to a "proper" coordinate system, the truth value of that statement is *uniquely* determined by an appropriate foliation of spacelike hypersurfaces. In this way, Hilbert unites the objectivity achieved by general covariance with the subjectivity of our experience of the world as causal, dissolving the appearance of conflict.

In sum, Hilbert's contribution to the Cauchy problem arises from the mathematical face of his causality problem. The problem is a genuine technical challenge facing generally covariant physics, and Hilbert's proposal for a solution remains familiar in the practice of contemporary General Relativity. For Hilbert, this mathematical problem is embedded in a deeper epistemological problem: that of reconciling our experience of a causally ordered and univocally determinate world with the four-dimensional structure of generally covariant physics. This problem has not gone away, and though Hilbert's proposal may be unfamiliar in contemporary discussions, we submit that it may be worth revisiting in earnest.⁸

⁷See, once again, Ohanian (2008, n. 56), cited above.

⁸Indeed, to go further, Hilbert's epistemological analysis of the differing status that should be accorded to general covariance versus causality might perhaps be suggestive to those working on the interpretation of General Relativity as a gauge theory, and the associated "problem of time" in quantum gravity.

3.7 Hilbert and Einstein Compared

Einstein abandoned general covariance for over two years, justifying this in part by appeal to his hole argument, and the “causality problem” that this argument poses. Then, with his generally covariant field equations in hand, Einstein restored causality by means of his “point coincidence argument”, which is best understood as asserting that systems of spacetime coordinates have no metrical or other physical significance, but serve merely as arbitrary labels for spacetime points.

We maintain that Hilbert’s “causality problem” was never that posed by Einstein’s hole argument⁹ despite superficial similarities in some of Hilbert’s discussions (see, for example, Hilbert 1916). In our opinion, the clearest way to see this is to recognize that Einstein’s solution to his own problem, concerning the status of spacetime coordinates, is something that Hilbert emphasized from the outset, and which furthermore does nothing to address the problem that Hilbert was addressing.

On the first point, it is significant that in 1915 Hilbert termed the labels for points in his four-dimensional spacetime “world parameters”. This terminology highlights the analogy with the arbitrary character of curve parameterizations in the calculus of variations. As Howard and Norton (1993) point out, the Göttingen mathematical community was thoroughly familiar with the use of arbitrary coordinates in the work of Lagrange, Gauss, and Riemann, and it seems highly unlikely, to say the least, that Hilbert was confused about this issue. This evidence is circumstantial, but compelling, and strongly supported by the statement of the problem that Hilbert gives in his Theorem 1 (Hilbert 1915a,b), where Hilbert makes clear that his “causality problem” arises due to *dependencies* among the field equations. In other words, Hilbert *started* from a position in which coordinates are merely arbitrary labels for spacetime points, and thus his “causality problem” cannot be that which Einstein expresses in the hole argument; rather, it arises when we seek to recover the mathematical property of Cauchy determination.

That Einstein’s and Hilbert’s problems with causality differ is further emphasized by the second point noted above: Einstein’s solution to *his* causality problem (via the point coincidence argument) does nothing to address Hilbert’s. Recall that, essentially, Einstein’s solution in his point coincidence argument posits a four-dimensional arrangement of events on a spacetime manifold. This four-dimensional arrangement does not come causally ordered, nor does it tell us how to recover univocal determination from our field equations. Both of these can be achieved by picking an appropriate co-ordinate system, and this is exactly what Hilbert points

⁹Stachel (1992) writes that while Einstein was “always a bit vague about just what he meant by causality” in his hole argument, Hilbert on the other hand “gave a quite precise meaning to the concept”, formulating it in the context of whether the field equations can be expressed in Cauchy normal form. Surely it is right that Hilbert was led to think about causality in the context of general covariance by Einstein’s concerns in the summer of 1915, and Stachel is of course exactly right that Hilbert’s version of the problem is stated in the precise mathematical language of the Cauchy problem. What we wish to emphasize is that the problem Hilbert thus arrives at is importantly *different* from that with which Einstein wrestled in his hole argument.

out that we need to do: in order to recover—via the field equations—our experience of the four-dimensional world of point events as univocally causally ordered, we need to make use of “proper” coordinate systems.¹⁰ This does *not* imply that Hilbert attributed physical significance to coordinate systems. On the contrary, as we have emphasized, and as Stachel (1992, p. 412) remarks, Hilbert assumed from the outset that “all solutions related by a coordinate transformation must be regarded as physically equivalent.”

3.8 Conclusion

Hilbert’s “causality problem” is not that which Einstein expressed in his “hole argument” and resolved with his “point coincidence argument”. Rather, Hilbert *begins* from the key premise of the “point coincidence argument”, that coordinates are arbitrary labels, and seeks to solve the mathematical and epistemological problems that then arise. The mathematical problem, and Hilbert’s proposed solution, remain standard fare in General Relativity today: the imposition of coordinate conditions for arriving at a solution to the Einstein Field Equations. The epistemological problem, of how we reconcile general covariance with our experience of the world as spatiotemporally and causally ordered and causally determinate, remains a puzzle admitting of no uncontroversial solution. Hilbert’s proposal piggy-backs on his solution to the mathematical problem and, in our opinion, its implications for the interpretation of General Relativity and associated epistemological issues deserve further exploration.¹¹

References

- Barbour, J. B. (2005). *Absolute or relative motion? Vol. 2: The deep structure of general relativity*. Oxford: Oxford University Press.
- Brading, K. A., & Ryckman, T. A. (2008). Hilbert’s ‘Foundations of Physics’: Gravitation and electromagnetism within the axiomatic method. *Studies in History and Philosophy of Modern Physics*, 39, 102–153.
- Brading, K. A., & Ryckman, T. A. (2009). Hilbert’s Axiomatic Method and his “Foundations of Physics”: Reconciling causality with the axiom of general invariance. In (Lehner et al. 2012, 175–199).
- Corry, L., Renn, J., & Stachel, J. (1997). Belated decision in the Hilbert-Einstein priority dispute. *Science*, 278, 1270–1273.
- Earman, J., Janssen, M., & Norton, J. (Eds.). (1993). *The attraction of gravitation: New studies in the history of general relativity* (Einstein studies vol. 5, pp. 30–62). Boston: Birkhäuser.

¹⁰See section on Hilbert’s causality problem, above.

¹¹While Hilbert himself addressed the epistemological problem within a Kantian framework (see Brading and Ryckman (2008, 2009)), it is not obvious that the core proposal requires this.

- Einstein, A. (1916). Die Grundlage der allgemeinen Relativitätstheorie. *Annalen der Physik*, 49, 769–822. Reprinted in (Einstein 1996, 284–337). Translated by W. Parrett and G. Jeffrey as “The Foundation of the General Theory of Relativity”, in H. Lorentz et al. (1923), 109–65.
- Einstein, A. (1995). *The collected papers of Albert Einstein. Vol. 4: The swiss years, writings, 1912–1914*, M. J. Klein, A. J. Kox, & R. Schulmann (Eds.). Princeton: Princeton University Press.
- Einstein, A. (1996). *The collected papers of Albert Einstein. Vol. 6: The Berlin years, writings, 1914–1917*, A. J. Kox, M. J. Klein, & R. Schulmann (Eds.). Princeton: Princeton University Press.
- Einstein, A., & Grossmann, M. (1914). Entwurf einer verallgemeinerten Relativitätstheorie und einer Theorie der Gravitation. *Zeitschrift für Mathematik und Physik*, 62, 225–259. Reprinted in (Einstein 1995, 302–344).
- Eisenstaedt, J., & Kox, A. J. (Eds.). (1992). *Studies in the history of general relativity* (Einstein studies, vol. 3). Boston: Birkhäuser.
- Hilbert, D. (1915a). Die Grundlagen der Physik (Erste Mitteilung). Annotated “Erste Korrektur meiner erste Note”, printer’s stamp date “6 Dez. 1915”. 13 pages with omissions. Published in (Sauer and Majer 2009, 317–330). Translation as “The foundations of physics (Proofs of first communication)”, in (Renn and Schemmel 2007, 989–1001).
- Hilbert, D. (1915b). Die Grundlagen der Physik (Erste Mitteilung). In *Nachrichten. Königliche Gesellschaft der Wissenschaften zu Göttingen. Mathematische-Physikalische Klasse* (pp. 395–407). Reprinted in (Sauer and Majer 2009, 28–46). Translation as “The foundations of physics (first communication)”, in (Renn and Schemmel 2007, 1003–1015).
- Hilbert, D. (1916). Das Kausalitätsprinzip in der Physik. Typescript of lectures, dated 21 and 28 November 1916. Bibliothek des Mathematisches Institut, Universität Göttingen. 17 pages. Published in (Sauer and Majer 2009, 335–346).
- Hilbert, D. (1917). Die Grundlagen der Physik (Zweite Mitteilung). *Nachrichten. Königliche Gesellschaft der Wissenschaften zu Göttingen. Mathematische-Physikalische Klasse* (pp. 53–76). Reprinted in (Sauer and Majer 2009, 47–72). Translation as “The foundations of physics (second communication)”, in (Renn and Schemmel 2007, 1017–1038).
- Hilbert, D. (1919/1992). *Natur und mathematisches Erkennen. Vorlesung gehalten 1919–1920 in Göttingen*. D. E. Rowe, Hrsg. Basel: Birkhäuser.
- Howard, D., & Norton, J. (1993). Out of the labyrinth? Einstein, Hertz, and the Göttingen answer to the hole argument. In (Earman, Janssen and Norton 1993, 30–62).
- Howard, D., & Stachel, J. (Eds.). (1989). *Einstein and the history of general relativity* (Einstein studies, vol. 1). Basel: Birkhäuser.
- Janssen, M., & Renn, J. (2007). Untying the knot: How Einstein found his way back to field equations discarded in the Zurich notebook. In M. Janssen, J. D. Norton, J. Renn, T. Sauer, & J. Stachel (Eds.), *The genesis of general relativity* (vol. 2, pp. 839–925). Einstein’s Zurich notebook: Commentary and essays. Dordrecht: Springer.
- Lehner, C., Renn, J., & Schemmel, M. (Eds.). (2012). *Einstein and the changing worldviews of physics* (Einstein studies, vol. 12). New York: Springer.
- Majer, U. (1995). Geometry, intuition and experience: From Kant to Husserl. *Erkenntnis*, 42, 261–285.
- Norton, J. D. (1984). How Einstein found his field equations: 1912–1915. *Historical Studies in the Physical Sciences*, 14, 253–316. Reprinted in (Howard and Stachel 1989, 101–159)
- Norton, J. D. (1993). General covariance and the foundations of general relativity: Eight decades of dispute. *Reports on Progress in Physics*, 56, 791–858.
- Ohanian, H. C. (2008). *Einstein’s mistakes*. New York: Norton.
- Renn, J., & Schemmel, M. (2007). *The genesis of general relativity*. Gravitation in the twilight of classical physics: The promise of mathematics. Dordrecht: Springer.
- Renn, J., & Stachel, J. (1999). Hilbert’s foundation of physics: From a theory of everything to a constituent of general relativity. Berlin: Max-Planck-Institut für Wissenschaftsgeschichte, Preprint 118. Reprinted in (Renn and Schemmel 2007, 858–973).
- Ryckman, T. A. (2005). *The reign of relativity*. Oxford: Oxford University Press.

- Sauer, T. (1999). The relativity of discovery: Hilbert's first note on the foundations of physics. *Archive for History of Exact Sciences*, 53, 529–575.
- Sauer, T. (2005). Einstein equations and Hilbert action: What is missing on page 8 of the proofs for Hilbert's first communication on the foundations of physics? *Archive for History of Exact Sciences*, 59, 577–590.
- Sauer, T., & U. Majer (Eds.). (2009). *David Hilbert's Lectures on the Foundations of Physics, 1915–1927*. Berlin: Springer.
- Stachel, J. (1989). Einstein's search for general covariance, 1912–1915. In (Howard and Stachel 1989, 63–100). Based on a paper circulating privately since 1980.
- Stachel, J. (1992). The Cauchy problem in general relativity - The early years. In (Eisenstaedt and Kox 1992, 407–418).

Part II
Testing General Relativity and Rival
Theories

Chapter 4

Putting General Relativity to the Test: Twentieth-Century Highlights and Twenty-First-Century Prospects



Clifford M. Will

4.1 Introduction

During the late 1960s, it was frequently said that “the field of general relativity is a theorist’s paradise and an experimentalist’s purgatory.” The field was not without experiments, of course: Irwin Shapiro, then at MIT, had just measured the relativistic retardation of radar waves passing the Sun (an effect that now bears his name); Robert Dicke of Princeton was claiming that the Sun was flattened by rotation in an amount whose effects on Mercury’s perihelion advance would put general relativity in jeopardy, and Joseph Weber of the University of Maryland was busy building gravitational wave antennas out of massive aluminum cylinders. Nevertheless the field was dominated by theory and by theorists. The field *circa* 1970 seemed to reflect Einstein’s own attitudes: although he was not ignorant of experiment, and indeed had a keen insight into the workings of the physical world, he felt that the bottom line was the *theory*. As he once famously said, if experiment were to contradict the theory, he would have “felt sorry for the dear Lord.”

Since that time the field has been completely transformed, and today experiment is a central component of gravitational physics and in some aspects is setting the agenda for the field. The breadth of current experiments, ranging from tests of classic general relativistic effects, to tests using gravitational waves, to ideas for testing the theory in astrophysical settings or on cosmic scales, attest to the ongoing vigor of experimental gravitation.

The great progress in testing general relativity during the latter part of the twentieth century featured three main themes:

C. M. Will (✉)

Department of Physics, University of Florida, Gainesville, FL, USA

e-mail: cmw@phys.ufl.edu

© Springer Science+Business Media, LLC, part of Springer Nature 2018

D. E. Rowe et al. (eds.), *Beyond Einstein*, Einstein Studies 14,

https://doi.org/10.1007/978-1-4939-7708-6_4

- The use of advanced technology. This included the high-precision technology associated with atomic clocks, laser and radar ranging, cryogenics, and delicate laboratory sensors, as well as access to space.
- The development of general theoretical frameworks. These frameworks allowed one to think beyond the narrow confines of general relativity itself, to analyze broad classes of theories, to propose new experimental tests, and to interpret the tests in an unbiased manner.
- The synergy between theory and experiment. To illustrate this, one needs only to note that the LIGO-Virgo Scientific Collaboration, engaged in one of the most important general relativity investigations – the detection of gravitational radiation – consists of over 1000 scientists. This is big science, reminiscent of high-energy physics, not general relativity!

Today, because of its elegance and simplicity, and because of its empirical success, general relativity is the standard model for our understanding of the gravitational interaction. Yet developments in particle theory and observations in cosmology suggest that it is probably not the entire story and that modifications of the basic theory may be required at some level. String theory generally predicts a proliferation of additional fields that could result in alterations of general relativity similar to that of the Brans-Dicke theory of the 1960s. In the presence of extra dimensions, the gravity that we feel on our four-dimensional “brane” of a higher-dimensional world could be somewhat different from a pure four-dimensional general relativity. And the observation that the expansion of the universe is accelerating has opened the possibility that modifications of general relativity on the largest scales might be required. However, any theoretical speculation along these lines *must* abide by the best current empirical bounds. Still, most of the current tests involve the weak-field, slow-motion limit of gravitational theory.

Putting general relativity to the test during the twenty-first century is likely to involve three main themes:

- Tests of strong-field gravity. These are tests of the nature of gravity near black holes and neutron stars, far from the weak-field regime of the solar system.
- Tests using gravitational waves. The detection of gravitational waves in 2015 has initiated a new form of astronomy, but it has also provided new tests of general relativity in the strong-field, radiative regime.
- Tests of gravity at extreme scales. The detected acceleration of the universe, the observed large-scale effects of dark matter, and the possibility of extra dimensions with effects on small scales have revealed how little is known about gravity on the largest and smallest scales.

In this paper we will review selected highlights of testing general relativity during the twentieth century and will discuss the potential for new tests in the twenty-first century. We begin in Section 4.2.1 with the “Einstein Equivalence Principle,” which underlies the idea that gravity and curved spacetime are synonymous, and describe its empirical support. Section 4.2.2 describes solar system tests of gravity in terms of experimental bounds on a set of “parametrized post-Newtonian” (PPN) parameters.

In Section 4.2.3 we discuss tests of general relativity using binary pulsar systems. Section 4.3.1 describes tests of gravitational theory that can be carried out using observations of gravitational radiation, and Section 4.3.2 describes the possibility of performing strong-field tests of general relativity. Tests of gravity at cosmological and submillimeter scales are significant topics in their own right and are beyond the scope of this paper. Concluding remarks are made in Section 4.4. For further discussion of topics in this paper, and for references to the primary literature, the reader is referred to *Theory and Experiment in Gravitational Physics* (Will 1993) and to the “living” review articles by Will (2014), Stairs (2003), Psaltis (2008), Mattingly (2005), Yunes and Siemens (2013), and Gair et al. (2013).

4.2 Twentieth-Century Highlights

4.2.1 *The Einstein Equivalence Principle*

The Einstein equivalence principle (EEP) is a powerful and far-reaching principle, which states that (i) test bodies fall with the same acceleration independently of their internal structure or composition (weak equivalence principle or WEP); (ii) the outcome of any local non-gravitational experiment is independent of the velocity of the freely falling reference frame in which it is performed (local Lorentz invariance or LLI); and (iii) the outcome of any local non-gravitational experiment is independent of where and when in the universe it is performed (local position invariance or LPI).

The Einstein equivalence principle is central to gravitational theory, for it is possible to argue convincingly that if EEP is valid, then gravitation must be described by “metric theories of gravity,” which state that (i) spacetime is endowed with a symmetric metric, (ii) the trajectories of freely falling bodies are geodesics of that metric, and (iii) in local freely falling reference frames, the non-gravitational laws of physics are those written in the language of special relativity.

General relativity is a metric theory of gravity but so are many others, including the Brans-Dicke theory.

To illustrate the high precisions achieved in testing EEP, we shall review tests of the weak equivalence principle, where one compares the acceleration of two laboratory-sized bodies of different composition in an external gravitational field. A measurement or limit on the fractional difference in acceleration between two bodies yields a quantity $\eta \equiv 2|a_1 - a_2|/|a_1 + a_2|$, called the “Eötvös ratio,” named in honor of Baron von Eötvös, the Hungarian physicist whose experiments carried out with torsion balances at the end of the nineteenth century were the first high-precision tests of WEP (see Figure 4.1). Later classic experiments by Dicke and Braginsky in the 1960s and 1970s improved the bounds by several orders of magnitude. Additional experiments were carried out during the 1980s as part of a search for a putative “fifth force” that was motivated in part by a reanalysis

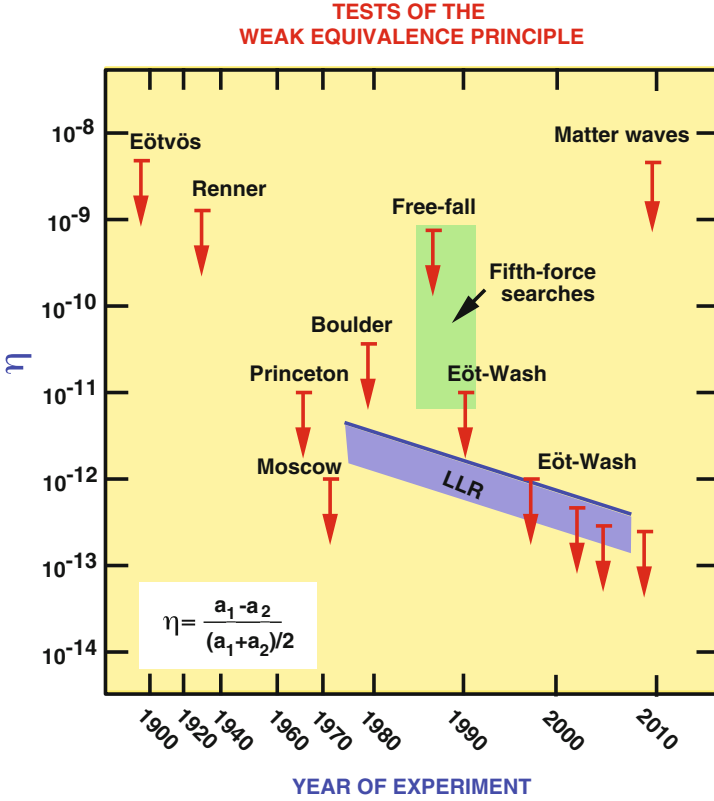


Fig. 4.1 Selected tests of the weak equivalence principle, showing bounds on the fractional difference in acceleration of different materials or bodies. Blue line and shading show evolving bounds on WEP for the Earth and the Moon from lunar laser ranging (LLR).

of Eötvös’ original data (the range of bounds achieved during that period is shown schematically in the region labeled “fifth force” in Figure 4.1).

The best limit on η currently comes from the “Eöt-Wash” experiments carried out at the University of Washington, which used a sophisticated torsion balance tray to compare the accelerations of bodies of different compositions toward the Earth, the Sun, and the galaxy. Another strong bound comes from lunar laser ranging (LLR), which checks the equality of free fall of the Earth and Moon toward the Sun. The results from laboratory and LLR experiments are:

$$\eta_{\text{Eöt-Wash}} < 3 \times 10^{-13}, \quad \eta_{\text{LLR}} < 3 \times 10^{-13}. \tag{4.1}$$

In fact, by using laboratory materials whose composition mimics that of the Earth and Moon, the Eöt-Wash experiments permit one to infer an unambiguous bound

from lunar laser ranging on the universality of acceleration of gravitational binding energy at the level of 9×10^{-4} (test of the Nordtvedt effect – see Section 4.2.2 and Table 4.1.)

High-precision WEP experiments can test superstring inspired models of scalar-tensor gravity or theories with varying fundamental constants in which weak violations of WEP can occur via nonmetric couplings. The project MICROSCOPE, designed to test WEP to a part in 10^{15} , was launched by the French space agency CNES in 2016. Other concepts for future improvements include advanced space experiments, experiments on suborbital rockets, lunar laser ranging, binary pulsar observations, and experiments with antihydrogen. For an update on past and future tests of WEP, see the series of articles introduced by Speake and Will (2012).

Very stringent constraints on local Lorentz invariance have been placed, notably by experiments that exploited laser-cooled trapped atoms to search for variations in the relative frequencies of different types of atoms as the Earth rotates around our velocity vector relative to the mean rest frame of the universe (as determined by the cosmic background radiation). For reviews, see Mattingly (2005), Will (2006), and Liberati (2013). Local position invariance has also been tested by gravitational redshift experiments and by tests of variations with cosmic time of fundamental constants. For a review of such tests, see Uzan (2011).

4.2.2 Solar System Tests

It was once customary to discuss experimental tests of general relativity in terms of the “three classical tests,” the gravitational redshift, which is really a test of the EEP, not of general relativity itself; the perihelion advance of Mercury, the first success of the theory; and the deflection of light, whose measurement in 1919 made Einstein a celebrity. However, the proliferation of additional tests as well as of well-motivated alternative metric theories of gravity made it desirable to develop a more general theoretical framework for analyzing both experiments and theories.

This “parametrized post-Newtonian (PPN) framework” dates back to Eddington in 1922 but was fully developed by Nordtvedt and Will in the period 1968–1972. When we confine attention to metric theories of gravity and further focus on the slow-motion, weak-field limit appropriate to the solar system and similar systems, it turns out that, in a broad class of metric theories, only the values of a set of numerical coefficients in the expression for the spacetime metric vary from theory to theory. The framework contains ten PPN parameters: γ , related to the amount of spatial curvature generated by mass; β , related to the degree of nonlinearity in the gravitational field; ξ , α_1 , α_2 , and α_3 , which determine whether the theory violates local position invariance or local Lorentz invariance in *gravitational* experiments (violations of the strong equivalence principle); and ζ_1 , ζ_2 , ζ_3 , and ζ_4 , which describe whether the theory has appropriate momentum conservation laws. In general relativity, $\gamma = 1$, $\beta = 1$, and the remaining parameters all vanish. For a complete exposition of the PPN framework, see Will (1993).

To illustrate the use of these PPN parameters in experimental tests, we cite the deflection of light by the Sun, an experiment that made Einstein an international celebrity when the sensational news of the Eddington-Clark eclipse measurements was relayed in November 1919 to a war-weary world. For a light ray which passes a distance d from the Sun, the deflection is given by

$$\begin{aligned}\Delta\theta &= \left(\frac{1+\gamma}{2}\right) \frac{4GM}{dc^2} \\ &= \left(\frac{1+\gamma}{2}\right) \times 1.7505 \left(\frac{R}{d}\right) \text{ arcsec},\end{aligned}\tag{4.2}$$

where M and R are the mass and radius of the Sun and G and c are the Newtonian gravitational constant and the speed of light. The “1/2” part of the coefficient can be derived by considering the Newtonian deflection of a particle passing by the Sun, in the limit where the particle’s velocity approaches c ; this was first calculated independently by Henry Cavendish around 1784 and Johann von Soldner around 1803. The second “ $\gamma/2$ ” part comes from the bending of “straight” lines near the Sun relative to lines far from the Sun, as a consequence of space curvature. A related effect, called the Shapiro time delay, an excess delay in travel time for light signals passing by the Sun, also depends on the coefficient $(1 + \gamma)/2$.

To illustrate the dramatic progress of experimental gravity since the dawn of Einstein’s theory, Figure 4.2 shows a history of results for $(1 + \gamma)/2$. “Optical” denotes measurements using visible light, made mainly during solar eclipses, beginning with the 1919 measurements of Eddington and his colleagues. Arrows denote values well off the chart from one of the 1919 eclipse expeditions and from others through 1947. “Radio” denotes interferometric measurements of radio-wave deflection, and “VLBI” denotes very long baseline radio interferometry, culminating in a global analysis of VLBI data on over 540 quasars and compact radio galaxies distributed over the entire sky, which verified GR at the 0.02 percent level. “Hipparcos” denotes the European optical astrometry satellite. “Galactic lensing” denotes a 2006 measurement of γ using stellar velocity-dispersion measurements and gravitational lensing data on 15 elliptical galaxies taken from the Sloan Digital Sky Survey. The GAIA astronomical observatory, launched in 2013, is a high-precision astrometric telescope (a successor to Hipparcos), which could, among other science goals, measure light deflection and γ to the 10^{-6} level.

Shapiro time delay measurements began in the late 1960s, by bouncing radar signals off Venus and Mercury; the most recent test used tracking data from the *Cassini* spacecraft on its way to Saturn, yielding a result at the 0.001 percent level.

Other experimental bounds on the PPN parameters, all consistent with general relativity, came from measurements of the perihelion shift of Mercury, bounds on the “Nordtvedt effect” (a possible violation of the weak equivalence principle for self-gravitating bodies) via lunar laser ranging, and pulsar measurements. Table 4.1 summarizes the current bounds.

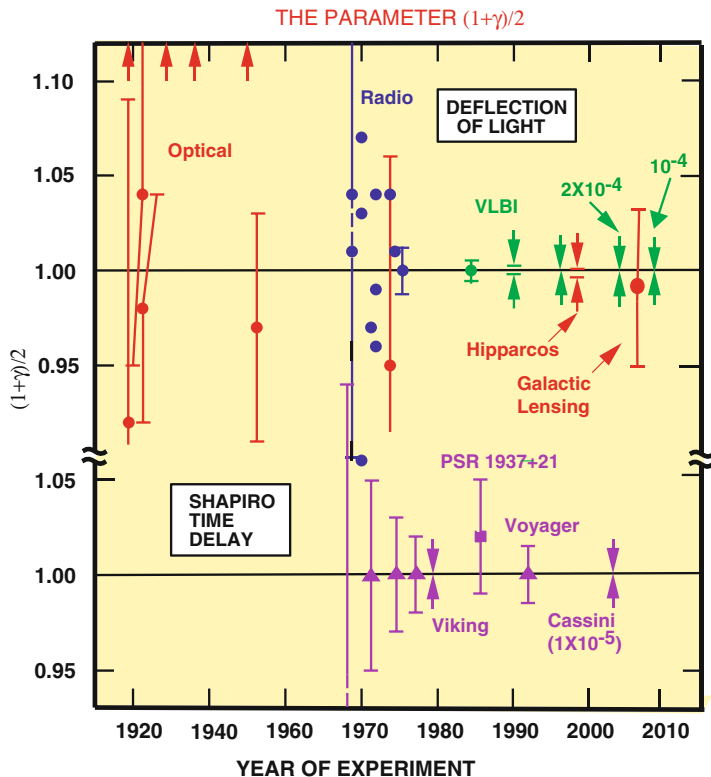


Fig. 4.2 Measurements of the coefficient $(1+\gamma)/2$ from observations of the deflection of light and of the Shapiro delay in propagation of radio signals near the Sun. The general relativity prediction is unity.

The perihelion advance of Mercury, the first of Einstein’s successes, is now known to agree with observation to better than a part in 10^4 , largely from improved data on the planet’s orbit provided by the Mercury Messenger orbiter. Although there was controversy during the 1960s about this test because of Dicke’s claims of an excess solar oblateness $J_{2\odot}$, which would result in an unacceptably large Newtonian contribution to the perihelion advance, it is now known from helioseismology that the oblateness is 2.2×10^{-7} , as expected from standard solar models, and too small to affect Mercury’s orbit, within the experimental error.

Scalar-tensor theories of gravity are characterized by a coupling function $\omega(\phi)$ whose size is inversely related to the “strength” of the scalar field relative to the metric. In the solar system, the parameter $|\gamma - 1|$, for example, is equal to $1/(2 + \omega(\phi_0))$, where ϕ_0 is the value of the scalar field today outside the solar system. Solar system experiments (primarily the Cassini results) constrain $\omega(\phi_0) > 40000$.

Future space missions could provide improved measurements of PPN parameters. BepiColombo, a Mercury orbiter mission planned for launch around 2018,

Table 4.1 Current limits on the PPN parameters

Parameter	Effect	Limit	Remarks
$\gamma - 1$	(i) time delay	2.3×10^{-5}	Cassini tracking
	(ii) light deflection	2×10^{-4}	VLBI
$\beta - 1$	(i) perihelion shift	8×10^{-5}	$J_{2\odot} = (2.2 \pm 0.1) \times 10^{-7}$
	(ii) Nordtvedt effect	2.3×10^{-4}	$\eta = 4\beta - \gamma - 3$ assumed
ξ	spin precession	4×10^{-9}	millisecond pulsars
α_1	orbital polarization	10^{-4}	lunar laser ranging
		4×10^{-5}	PSR J1738+0333
α_2	spin precession	2×10^{-9}	millisecond pulsars
α_3	pulsar acceleration	2×10^{-20}	pulsar \dot{P} statistics
ζ_1	–	2×10^{-2}	combined PPN bounds
ζ_2	binary acceleration	4×10^{-5}	\ddot{P}_p for PSR 1913+16
ζ_3	Newton's 3rd law	10^{-8}	lunar acceleration
ζ_4	–	–	not independent

could, among other measurements, determine J_2 of the Sun to 10^{-8} and improve bounds on a time variation of the gravitational constant. The Laser Astrometric Test of Relativity (LATOR) and the Astrodynamical Space Test of Relativity using Optical Devices (ASTROD), involving laser ranging to one or more satellites on the far side of the Sun, could measure γ to a part in 10^8 and could possibly detect second-order effects in light propagation. The Apache Point Observatory for Lunar Laser-ranging Operation (APOLLO) project, a joint effort by researchers from the Universities of Washington, Seattle, and California, San Diego, is using enhanced laser and telescope technology, together with a good, high-altitude site in New Mexico, with the goal of improving the lunar laser-ranging bound by as much as an order of magnitude.

The NASA mission called Gravity Probe B completed its measurement of the Lense-Thirring and geodetic precessions of gyroscopes in Earth orbit. Launched on April 20, 2004, for a 16-month mission, it consisted of four spherical fused quartz rotors coated with a thin layer of superconducting niobium, spinning at 70–100 Hz, in a spacecraft containing a telescope continuously pointed toward a distant guide star (IM Pegasi). Superconducting current loops encircling each rotor measured the change in direction of the rotors by detecting the change in magnetic flux through the loop generated by the London magnetic moment of the spinning superconducting film. The spacecraft orbited the Earth in a polar orbit at 650 km altitude. In 2011, GPB reported a 20 percent measurement of the 41 milliarcsecond per year frame dragging or Lense-Thirring effect caused by the rotation of the Earth and a 0.3 percent measurement of the larger 6.6 arcsecond per year geodetic precession caused by space curvature (Everitt et al. 2011).

A complementary test of the Lense-Thirring precession involved measuring the precession of the orbital planes of two Earth-orbiting laser-ranged satellites called LAGEOS, using up-to-date models of the gravitational field of the Earth in order to subtract the dominant Newtonian precession with sufficient accuracy to yield a measurement of the relativistic effect, good to about 10 percent. Including data from a third laser-ranged satellite called LARES, launched in 2012, could improve the measurement, possibly to the one percent level.

As one final illustration of the precision and depth of the constraints on alternative theories of gravity, we mention the strange case of Alfred North Whitehead's 1922 theory of gravity. Uncomfortable with the fact that, in general relativity, the causal relationships in spacetime are not known *a priori* until the field equations have been solved, Whitehead proposed a Lorentz invariant theory of gravity in which the physical metric was constructed algebraically from a flat background Minkowski metric and matter variables (mass and velocities). The theory had the unusual property that, for a point mass at rest, the metric was mathematically equivalent to the Schwarzschild metric. As a consequence, Whitehead's theory could not be experimentally distinguished from general relativity using the standard tests of the day, such as light deflection, perihelion advance, or gravitational redshift. This led to a conundrum of how to select between competing theories that equally satisfy experimental observations.

But in 1971, we pointed out that when the theory was extended in a natural way to more than one gravitating body or to extended bodies, then the gravitational attraction between any pair of masses in the presence of a third body would be anisotropic, that is, dependent upon the orientation of the pair relative to the distant body (an effect additional to the normal tidal gravitational effects). This effective "anisotropy in Newton's constant G " would result in anomalous tide-like distortions of the Earth in the presence of the mass of the galaxy that were ruled out by precise measurements made with gravimeters. This however did not totally end the fascination with Whitehead's theory among some philosophically oriented scholars and so Gibbons and Will (2008) embarked on a serial killing of Whitehead's theory, pointing out that it actually fails the test of experiment in five different ways:

- *Anisotropy in G .* A reanalysis of Will's 1971 result verified that the theory violates gravimeter data on anomalous Earth tides by a factor of at least 100.
- *Nordtvedt effect.* The theory predicts a violation of the equivalence principle for gravitating bodies, leading to a Nordtvedt effect in lunar laser ranging 400 times larger than the data will permit.
- *Birkhoff's theorem and LAGEOS data.* It was already known in the 1950s that the theory predicts that the metric of a static, spherically symmetric *finite-sized* body has an additional size-dependent contribution. This contributes an additional advance of the perigee of the LAGEOS II satellite, in disagreement with observations by a factor of 10.
- *Momentum non-conservation.* The theory predicts an acceleration of the center of mass of a binary system, an effect ruled out by binary pulsar data by a factor of a million.

- *Gravitational radiation damping.* The theory predicts anti-damping in binary orbits due to gravitational radiation, in violation of binary pulsar data by four orders of magnitude.

The purpose of this list of failures is not to gang up on Whitehead but rather to illustrate that matching the Schwarzschild geometry is no longer sufficient to match experimental tests and that the current generation of empirical data strongly and deeply constrains the theoretical possibilities.

4.2.3 Binary Pulsars

The binary pulsar PSR 1913+16, discovered in 1974 by Joseph Taylor and Russell Hulse, provided important new tests of general relativity, specifically of gravitational radiation and of strong-field gravity. Through precise timing of the pulsar “clock,” the important orbital parameters of the system were measured with exquisite precision. These include nonrelativistic “Keplerian” parameters, such as the eccentricity e , and the orbital period (at a chosen epoch) P_b , as well as a set of relativistic “post-Keplerian” parameters (see Table 4.2). The first PK parameter, $\langle \dot{\omega} \rangle$, is the mean rate of advance of periastron, the analogue of Mercury’s perihelion shift. The second, denoted γ' , is the effect of special relativistic time dilation and the gravitational redshift on the observed phase or arrival time of pulses, resulting from the pulsar’s orbital motion and the gravitational potential of its companion. The third, \dot{P}_b , is the rate of decrease of the orbital period; this is taken to be the result of gravitational radiation damping (apart from a small correction due to galactic differential rotation). Two other parameters, s and r , are related to the Shapiro time

Table 4.2 Parameters of selected binary pulsars

Parameter	Symbol	Value ¹ in PSR1913+16	Value ¹ in J0737-3039
Keplerian parameters			
Eccentricity	e	0.6171334(5)	0.0877775(9)
Orbital period	P_b (day)	0.322997448911(4)	0.10225156248(5)
Post-Keplerian parameters			
Periastron advance	$\langle \dot{\omega} \rangle$ ($^{\circ}\text{yr}^{-1}$)	4.226598(5)	16.8995(7)
Redshift/time dilation	γ' (ms)	4.2992(8)	0.386(3)
Orbital period derivative	\dot{P}_b (10^{-12})	-2.423(1)	-1.25(2)
Shapiro delay (range)	r (μs)		6.2(3)
Shapiro delay ($\sin i$)	s		0.9997(4)

¹Numbers in parentheses denote errors in last. digit

delay of the pulsar signal if the orbital inclination is such that the radio signal from the pulsar passes the companion in close proximity; s is a direct measure of the orbital inclination $\sin i$. According to general relativity, the first three post-Keplerian effects depend only on e and P_b , which are known, and on the two stellar masses which are unknown. By combining the observations of PSR 1913+16 with the general relativity predictions, one obtains both a measurement of the two masses and a test of the theory, since the system is overdetermined. The results are

$$m_1 = 1.4398 \pm 0.0002 M_\odot, \quad m_2 = 1.3886 \pm 0.0002 M_\odot, \\ \dot{P}_b^{\text{GR}} / \dot{P}_b^{\text{OBS}} = 0.997 \pm 0.002. \quad (4.3)$$

The accuracy in measuring the relativistic damping of the orbital period is now limited by uncertainties in our knowledge of the relative acceleration between the solar system and the binary system as a result of galactic differential rotation.

The results also test the strong-field aspects of metric gravitation in the following way: the neutron stars that comprise the system have very strong internal gravity, which contributes as much as several tenths of the rest mass of the bodies (compared to the orbital energy, which is only 10^{-6} of the mass of the system). Yet in general relativity, the internal structure is “effaced” as a consequence of the strong equivalence principle (SEP), a stronger version of EEP that includes *gravitationally* bound bodies and local *gravitational* experiments. As a result, the orbital motion and gravitational radiation depend *only* on the masses m_1 and m_2 and not on their internal structure, apart from standard tidal and spin-coupling effects. By contrast, in alternative metric theories, SEP is not valid in general, and internal structure effects can lead to significantly different behaviors, such as the emission of dipole gravitational radiation. Unfortunately, in the case of scalar-tensor theories of gravity, because the neutron stars are so similar in PSR 1913+16 (and in other double-neutron star binary pulsar systems), dipole radiation is suppressed by symmetry; the best bound on the coupling parameter $\omega(\phi_0)$ from PSR 1913+16 is in the hundreds. By contrast, the close agreement of the data with the predictions of general relativity constitutes a kind of “null” test of the effacement of strong-field effects in that theory. On the other hand, strong bounds on dipole radiation have been placed using pulsars with white-dwarf companions, including J1738+0333 and J11416545.

The “double pulsar” J0737-3039, discovered in 2003, is a binary system with two detected pulsars, in a 147-minute orbit seen almost edge on, with eccentricity $e = 0.09$, and a periastron advance of 17° per year (the secondary pulsar has since undergone sufficient precession of its spin axis that it no longer beams a signal toward the Earth). A variety of novel tests of relativity, neutron star structure, and pulsar magnetospheric physics have been carried out in this system. And because of its relative proximity to the Earth and favorable location in the galaxy, measurement of the orbital period decrease will not be limited by galactic acceleration effects, and so this system will eventually surpass the Hulse-Taylor system in a precision test of gravitational-wave damping.

The remarkable triple system, J0337+1715, was reported in 2014. It consists of a 2.73 millisecond pulsar ($M = 1.44M_{\odot}$) with extremely good timing precision, accompanied by *two* white dwarfs in coplanar circular orbits. The inner white dwarf ($0.1975M_{\odot}$) has an orbital period of 1.629 days, with $e = 6.918 \times 10^{-4}$, and the outer white dwarf ($0.41M_{\odot}$) has a period of 327.26 days, with $e = 3.536 \times 10^{-2}$. This is an ideal system for testing the Nordtvedt effect in the strong-field regime. For reviews of binary pulsar tests, see Stairs (2003) and Damour (2009).

4.3 Twenty-First-Century Prospects

4.3.1 Gravitational-Wave Tests of Gravitation Theory

The 2015 detection of gravitational radiation by the ground-based laser interferometer observatory LIGO in the USA (Abbott et al. 2016) has ushered in a new era of gravitational-wave astronomy (for a pre-discovery review, see Hough and Rowan 2000). Over time, other advanced detectors – Virgo in Europe, KAGRA in Japan, and INDIGO in India – will join LIGO in a worldwide network of gravitational-wave observatories. The European Space Agency, with support from the NASA, Space Agency, plans to launch a space-based detector around 2030, nominally called LISA, to explore the low-frequency band of gravitational waves around a millihertz. And in the very low-frequency, nanohertz regime, coherent timing of millisecond pulsars distributed around the sky could also detect waves.

Furthermore, such observations promise to yield new and interesting tests of general relativity in its radiative regime. Indeed, the LIGO detection of waves from the inspiraling binary black hole system, dubbed GW150914, already yielded one such test.

This test concerns the speed of gravitational waves. According to general relativity, in the limit in which the wavelength of gravitational waves is small compared to the radius of curvature of the background spacetime, the waves propagate along null geodesics of the background spacetime, *i.e.*, they have the same speed, c , as light. In other theories, the speed could differ from c because of coupling of gravitation to “background” gravitational fields. For example, in some theories with a flat background metric η , gravitational waves follow null geodesics of η , while light follows null geodesics of g . In brane-world scenarios, the apparent speed of gravitational waves could differ from that of light if the former can propagate off the brane into the higher-dimensional “bulk.” Another way in which the speed of gravitational waves could differ from c is if gravitation were propagated by a massive field (a massive graviton), in which case v_g would depend on the wavelength λ of the gravitational waves according to $v_g/c \approx 1 - \lambda^2/2\lambda_g^2$, where $\lambda_g = h/m_g c$ is the graviton Compton wavelength ($\lambda_g \gg \lambda$ is assumed).

The most obvious way to test for a massive graviton is to compare the arrival times of a gravitational wave and an electromagnetic wave from the same event,

e.g., a supernova. For a source at a distance of hundreds of megaparsecs, and for a relative arrival time between the two signals of order seconds, the resulting bound would be of order $|1 - v_g/c| < 5 \times 10^{-17}$. A bound such as this cannot be obtained from solar system tests at current levels of precision.

However, there is a situation in which a bound on the graviton mass can be set using gravitational radiation alone, with no need for an electromagnetic counterpart. That is the case of the inspiraling compact binary, the final stage of evolution of systems like the binary pulsar, in which the loss of energy to gravitational waves has brought the binary to an inexorable spiral toward a final merger. Because the frequency of the gravitational radiation sweeps from low frequency at the initial moment of observation to higher frequency at the final moment, the speed of the gravitational waves emitted will vary, from lower speeds initially to higher speeds (closer to c) at the end. This will cause a distortion of the observed phasing of the waves as a function of wavelength. Furthermore, through the technique of matched filtering, whereby a theoretical model of the wave and its phase evolution is cross-correlated against the detected signal, the parameters of the compact binary can be measured accurately, along with the parameter, the graviton Compton wavelength, that governs the distortion.

This analysis was performed on the signal from GW150914, the merger of two black holes, yielding a lower bound on λ_g of the order of several times 10^{13} km, slightly better than the bound $\lambda_g > 2.8 \times 10^{12}$ km, derived from solar system dynamics, which limits the presence of a Yukawa modification of Newtonian gravity of the form $V(r) = (GM/r)e(-r/\lambda_g)$. The space-based LISA antenna could place a bound on λ_g on the order of 10^{16} km.

Another test that will be possible involves verifying the polarization content of the waves; general relativity predicts only two polarization modes out of a possible six, irrespective of the source.

4.3.2 Tests of Gravity in the Strong-Field Regime

One of the main difficulties of testing GR in the strong-field regime is the possibility of contamination by uncertain or complex physics. In the solar system, weak-field gravitational effects can in most cases be measured cleanly and separately from non-gravitational effects. The remarkable cleanliness of most binary pulsars permitted precise measurements of gravitational phenomena in a strong-field context. Unfortunately, nature is rarely so kind. Still, under suitable conditions, qualitative and even quantitative strong-field tests of GR could be carried out.

One example is the exploration of the spacetime near black holes and neutron stars via accreting matter. Studies of certain kinds of accretion known as advection-dominated accretion flow (ADAF) in low-luminosity binary X-ray sources may yield the signature of the black hole event horizon. The spectrum of frequencies of quasiperiodic oscillations (QPO) from galactic black hole binaries may permit measurement of the spins of the black holes. Aspects of strong-field gravity and

frame dragging may be revealed in spectral shapes of iron fluorescence lines from the inner regions of accretion disks around black holes and neutron stars. Measurements of the detailed shape of the infrared image of the accretion flow around the 4 million solar-mass black hole SgrA* in the center of our galaxy or of the orbits of stars or pulsars in close proximity to the hole could also provide tests of the spacetime black hole metric. Because of uncertainties in the detailed models, the results to date of studies like these are suggestive at best, but the combination of higher-resolution observations and better modeling could lead to striking tests of strong-field predictions of GR. For a review of such tests, see Psaltis (2008).

The best tests of GR in the strong-field limit may come from gravitational-wave observations (see Yunes and Siemens (2013) and Gair et al. (2013) for reviews). The ground-based interferometers have detected and will continue to detect the gravitational waves from the final inspiral and merger of pairs of stellar-mass black holes or neutron stars. Comparison of the observed waveform with the predictions of GR from a combination of analytic and numerical techniques can test the theory in the most dynamical, strong-field limit. In fact a preliminary test of this kind was carried out using data from the *second* LIGO detection, GW151226. The space antenna LISA may observe as many as 100 mergers of massive black holes per year, with large signal-to-noise ratio. Such observations could provide precise measurements of black hole masses and spins and could test the “no hair” theorems of black holes by detecting the spectrum of quasi-normal “ringdown” modes emitted by the final black hole. Observations by LISA of the hundreds of thousands of gravitational-wave cycles emitted when a small black hole inspirals onto a massive black hole could test whether the geometry of a black hole actually corresponds to the “hair-free” Kerr metric.

4.4 Conclusions

Einstein’s general theory of relativity altered the course of science. It was a triumph of the imagination and of theory, but in the early years following its formulation, experiment played a secondary role. In the final four decades of the twentieth century, we witnessed a second triumph for Einstein, in the systematic, high-precision experimental verification of general relativity. It has passed every test with flying colors.

But the work is not done. During the twenty-first century, we may look forward to the possibility of tests of strong-field gravity in the vicinity of black holes and neutron stars. Electromagnetic and gravitational-wave astronomy will play a critical role in probing this largely unexplored aspect of general relativity.

General relativity is now the “standard model” of gravity. But as in particle physics, there may be a world beyond the standard model, beyond Einstein. Quantum gravity, strings, and branes may lead to testable effects beyond standard general relativity. Experimentalists and observers can be counted on to continue a vigorous search for such effects using laboratory experiments, particle accelerators,

gravitational-wave detectors, space telescopes, and cosmological observations, well into the twenty-first century.

Acknowledgements This work was supported in part by the US National Science Foundation, Grant No. PHY 06-52448, and by the National Aeronautics and Space Administration, Grant No. NNG-06GI60G. The 2016 update of the article was supported by NSF Grant Nos. PHY 13-06069 and 16-00188. We are grateful for the hospitality of the Institut d’Astrophysique de Paris, where the initial version of this paper was prepared.

References

- Abbott, B. P., Abbott, R., Abbott, T. D., Abernathy, M. R., Acernese, F., Ackley, K., et al. (The LIGO-Virgo Scientific Collaboration)(2016). Observation of gravitational waves from a binary black hole merger. *Physical Review Letters*, *116*, 061102.
- Damour, T. (2009). Binary systems as test-beds of gravity theories. In M. Colpi, P. Casella, V. Gorini, U. Moschella, & A. Possenti, *Physics of relativistic objects in compact binaries: From birth to coalescence: Astrophysics and space science library* (Vol. 359, p. 1). New York: Springer.
- Everitt, C. W. F., DeBra, D. B., Parkinson, B. W., Turneaura, J. P., Conklin, J. W., Heifetz, M. I., et al. (2011). Gravity probe B: Final results of a space experiment to test general relativity. *Physical Review Letters*, *106*, 221101.
- Gair, J. L., Vallisneri, M., Larson, S. L., & Baker, J. G. (2013). Testing general relativity with low-frequency, space-based gravitational-wave detectors. *Living Reviews in Relativity*, *16*, 7 [Online article]: Cited 15 Nov. 2016, <https://doi.org/10.12942/lrr-2013-7>.
- Gibbons, G., & Will, C. M. (2008). On the multiple deaths of Whitehead’s theory of gravity. *Studies in History and Philosophy of Modern Physics*, *39*, 41–61.
- Hough, J., & Rowan, S. (2000). Gravitational wave detection by interferometry (ground and space). *Living Reviews in Relativity*, *3*, 3 [On-line article]: Cited 15 Nov. 2016, <https://doi.org/10.12942/lrr-2000-3>.
- Liberati, S. (2013). Tests of Lorentz invariance: A 2013 update. *Classical Quantum Gravity*, *30*, 133001.
- Mattingly, D. (2005). Modern tests of Lorentz invariance. *Living Reviews in Relativity*, *8*, 5 [On-line article]: Cited 15 Nov. 2016, <https://doi.org/10.12942/lrr-2005-5>.
- Psaltis, D. (2008). Probes and tests of strong-field gravity with observations in the electromagnetic spectrum. *Living Reviews in Relativity*, *11*, 9 [On-line article]: Cited 15 Nov. 2016, <https://doi.org/10.12942/lrr-2008-9>.
- Speake, C. C., & Will, C. M. (2012). Tests of the weak equivalence principle. *Classical and Quantum Gravity*, *29*, 180301.
- Stairs, I. H. (2003). Testing general relativity with pulsar timing. *Living Reviews in Relativity*, *6*, 5 [On-line article]: Cited 15 Nov. 2016, <https://doi.org/10.12942/lrr-2003-5>.
- Uzan, J.-P. (2011). Varying constants, gravitation and cosmology. *Living Reviews in Relativity*, *14*, 2 [Online article]: Cited 15 Nov. 2016, <https://doi.org/10.12942/lrr-2011-2>.
- Will, C. M. (1993). *Theory and experiment in gravitational physics*. Cambridge: Cambridge University Press.
- Will, C. M. (2006). Special relativity: A centenary perspective. In T. Damour, O. Darrigol, B. Duplantier & V. Rivasseau (Eds.), *Einstein 1905-2005: Poincaré Seminar 2005* (p. 33). Basel: Birkhäuser.

- Will, C. M. (2014). The confrontation between general relativity and experiment. *Living Reviews in Relativity*, 17, 4 [Online article]: Cited 15 Nov. 2016, <https://doi.org/10.12942/lrr-2014-4>.
- Yunes, N., & Siemens, X. (2013). Gravitational-wave tests of general relativity with ground-based detectors and pulsar-timing arrays. *Living Reviews in Relativity*, 16, 9 [Online article]: Cited 15 Nov. 2016, <https://doi.org/10.12942/lrr-2013-9>.

Chapter 5

Rotating Hollow and Full Spheres: Einstein, Thirring, Lense, and Beyond



Herbert Pfister

5.1 Introduction

Rotating systems were of decisive importance in Einstein's struggle for a relativistic theory of gravity. Already in 1909 he wrote in a letter to Sommerfeld (Klein et al. 1993, Document 179): "The treatment of the uniformly rotating rigid body seems to me to be of great importance on account of an extension of the relativity principle to uniformly rotating systems . . ." In 1912, the consideration of the rigidly rotating disk promoted the idea that a relativistic gravity theory should be based on a curved space-time manifold (Stachel 1980). A short paper from 1912 (Einstein 1912c), although not treating rotating systems, is important for our topic because Einstein introduced here for the first time the model of an infinitely thin spherical mass shell and argued that a linear acceleration of this mass shell should induce a dragging of test masses inside the shell. In the Einstein–Besso manuscript (Klein et al. 1995, Document 14), and in Einstein's famous speech at the Congress of Natural Scientists and Physicians in Vienna (Einstein 1913), Einstein calculated, within the so-called Entwurf theory (Einstein and Grossmann 1913), the dragging of test masses inside a rotating mass shell and the motion of the nodes of planets or moons in the exterior

This contribution has been edited and updated by Markus King, University of Applied Sciences Albstadt-Sigmaringen, D-72458 Albstadt-Ebingen, Germany.

H. Pfister (✉) (deceased)

Institute of Theoretical Physics, University of Tübingen, D-72076 Tübingen, Germany

e-mail: rowe@mathematik.uni-mainz.de; tsauer@uni-mainz.de; scott.walter@univ-nantes.fr

© Springer Science+Business Media, LLC, part of Springer Nature 2018

D. E. Rowe et al. (eds.), *Beyond Einstein*, Einstein Studies 14,

https://doi.org/10.1007/978-1-4939-7708-6_5

field of a rotating central body, with results similar to those found in the later work of Thirring and Lense, based on general relativity.

Hans Thirring started calculations for rotating hollow and full spheres in the weak-field limit of general relativity in April 1917, but he confined himself to the diagonal metric components $g_{\mu\mu}$ whose nontrivial parts are of second order in the angular velocity ω and are considered by Thirring as centrifugal effects. Only after an exchange of letters with Einstein in July–August 1917, in which Einstein instructed Thirring that Coriolis effects of first order in ω , resulting from the off-diagonal metric components g_{ti} , are much bigger and nearer to observation than centrifugal effects, was Thirring able to calculate the Coriolis and “centrifugal” effects inside a rotating hollow sphere in the paper (Thirring 1918) and, with Josef Lense, to calculate the motion of the nodes of planets and moons in the paper (Lense and Thirring 1918).

Still, these papers, in particular the results of second order in ω in Thirring (1918), contain severe deficiencies which were clarified and corrected only gradually in later years by other authors. The proof that there exist appropriate rotating mass shells with correct Coriolis and centrifugal forces, and no other inertial forces in their interior, had even to wait until 1985 (Pfister and Braun 1985). Hereby, the demand made by Ernst Mach for “relativity of rotation”—which we understand in this paper according to the formulation, “Obviously it does not matter whether we consider the earth as rotating around its axis, or if we think of a static earth, and the celestial bodies rotating around it” in Mach (1872)—is optimally fulfilled for this model class in general relativity. The further conjecture that in general relativity every acceleration field can (at least in a finite space-time region) be understood as a gravitational field was formulated in Pfister and Braun (1985) as the “quasiglobal principle of equivalence.” In 2005 we succeeded in giving a partial proof of this conjecture by showing that also in a (first order) linearly accelerating mass shell “correct” dragging forces are induced (Pfister et al. 2005).

The models of isolated rotating mass shells, being embedded in an asymptotically flat space-time, have sometimes been criticized for not fulfilling cosmological boundary conditions for which Mach had formulated his conjecture of relativity of rotation. In the last years it was possible, however, to show that the dragging results for isolated rotating mass shells carry over, with only minor modifications, to rotationally disturbed Friedmann–Lemaître–Robertson–Walker (FLRW) cosmologies. Furthermore, it was shown that the Machian dragging effects in general relativity are produced by the (non-causal) constraints of the Einstein equations. Recent observations have provided increasing evidence of the fact that “relativity of rotation” is perfectly realized in our universe because the relative angular velocity between “local inertial systems” and the most distant galaxies and quasars is in some cases estimated to be below 10^{-9} of the earth’s angular velocity.

5.2 Einstein's 1913 Work on Rotating Spheres

Before coming to this work on rotating spheres, it is necessary to mention a four-page paper by Einstein from 1912 (Einstein 1912c), which introduced new concepts and ideas central to work on rotating spheres. One such new concept is the infinitely thin, spherical mass shell. This toy model is not only important for Einstein's early work on general relativity and the subsequent work of Thirring but has rendered an extremely useful service in GR until today for at least two reasons. Firstly, it constitutes the optimal substitute for the point mass of Newtonian mechanics, which is forbidden in general relativity due to the gravitational collapse phenomenon. Secondly, the shell model allows the calculation of mass effects in general relativity by solving only the vacuum Einstein equations inside and outside the shell, where the energy-momentum tensor of the shell is "automatically" given by the discontinuities between the normal derivatives of the interior and exterior solutions at the shell position.

With the help of such a mass shell (mass M , radius R), Einstein now postulated—on the basis of his preliminary scalar relativistic gravity theory (Einstein 1912a,b)—two new physical effects. For test masses m , he "derived" a mass increase $m \rightarrow m' = (1 + M/R)m$ (in units in which Newton's gravitational constant and the light velocity have the value 1) if these test masses are brought to the interior of the shell. Using this mass increase, Einstein calculated that a linear acceleration Γ of the whole mass shell induces an acceleration $\gamma = (3M/2R)\Gamma$ of the interior test masses. To my knowledge, this is the first concrete calculation of a gravitational dragging effect in the whole physics literature.

From today's perspective, and on the basis of general relativity, one should make the following comments concerning the paper (Einstein 1912c): A mass increase of test bodies due to the presence of a gravitational field had been controversially discussed for a long time until Carl Brans (Brans 1962) convincingly showed that this is only an untestable coordinate effect in general relativity. In principle, this could have been seen already in 1912 because it is hard to imagine how one should measure such a mass increase if gravity acts universally on all systems and measuring devices and can never be shielded. Linear dragging is a real effect in general relativity, but it is difficult to model this effect consistently (see Section 5.6), and the effect has not yet been experimentally tested. A critical remark should be made concerning the title, "Is there a gravitational action in analogy to electromagnetic induction?" of the paper (Einstein 1912c): Although gravitomagnetism is an established phenomenon within the tensorial theory, "general relativity," a gravitational action in analogy to the (vectorial!) electromagnetic induction is hardly imaginable in Einstein's scalar gravity theory (Einstein 1912a,b).

The Einstein–Besso manuscript (Klein et al. 1995, Document 14) and Einstein's Vienna speech (Einstein 1913) are based on the Entwurf theory (Einstein and Grossmann 1913), a tensorial, relativistic gravity theory based on a pseudo-Riemannian space-time which, however, is not yet generally covariant. The main issue of this manuscript is the perihelion advance of Mercury, for which, however,

a value of 18 arcsec/year was calculated, in contrast to the observational value of 43 arcsec/year which was, presumably, the reason why Einstein never published this work. However, the later parts of the Einstein–Besso manuscript contain other interesting and very elegantly calculated results:

- a) A Coriolis force inside a spherical rotating mass shell, leading to a dragging of test particles inside this shell, the ratio d between the angular velocity of the test particles and of the mass shell being $d = 2M/3R$, half the value which Thirring derived in Thirring (1918) in general relativity. This is the only part of the manuscript entering Einstein’s Vienna speech (Einstein 1913), where he also remarks that “unfortunately, the expected effect is so small that we cannot hope to verify it in terrestrial experiments or in astronomy.”
- b) A dragging of test particles inside a linearly accelerated mass shell: $\gamma = (2M/R)\Gamma$, a factor $4/3$ bigger than in the scalar theory (Einstein 1912c), and now derived without the dubious detour of a mass increase due to a gravitational field.
- c) A motion of the nodes of planets in the field of the rotating sun. If one compares their result with the later calculation in Lense and Thirring (1918) in general relativity and adjusts the different notation, it is seen that the effect in the Entwurf theory is only $1/4$ of the effect in general relativity.

5.3 The Contributions of Thirring and Lense from the Years 1917 to 1918 and Their Dependence on Einstein’s Intervention

A later article by Thirring (1966), a tribute on the 50th anniversary of Mach’s death, reveals that he originally tried to set up an experiment for measuring centrifugal effects inside a rotating hollow cylinder. This is somewhat strange because it should have been clear to Thirring that such a relativistic correction to Newton’s theory of gravity of second order in ω contains the factors M/R and $(v/c)^2$ and therefore is below the ridiculously small ratio 10^{-32} for all conceivable laboratory experiments ($M < 1000$ kg, $R > 1$ m, $v/c < 10^{-4}$).

The formation of the theoretical work of Thirring (and Lense) on rotating spheres can be very well traced in Thirring’s 156-page notebook, “Wirkung rotierender Massen” (Thirring 1917), from the years 1917 to 1922. (For a more detailed analysis of this notebook, see Pfister 2007.) The first third of the notebook contains calculations of metric components and Christoffel symbols for rotating hollow and full spheres in the weak-field limit $M \ll R$ of general relativity, but Thirring considers only the diagonal metric components $g_{\mu\mu}$ whose deviations from the Minkowski metric are of order ω^2 . Parts of the calculations are wrong, and some parts are crossed out by Thirring himself. For the date July 17, 1917, the notebook contains the draft of a letter to Einstein (Schulmann et al. 1998, Document 361) in which Thirring communicates the g_{44} component for the rotating mass shell,

mentions the surprising axial component of this “centrifugal force,” and asks Einstein whether he could think of an experimental confirmation of such a force for the innermost moon of Jupiter. Einstein’s answer of August 2, 1917 (Schulmann et al. 1998, Document 369) is quite short but it exposes the weak points in Thirring’s work (up to this time) in an admirably clear and concise way:

As to your example of the hollow sphere it has only to be said that besides the centrifugal field whose axial component you interpret so nicely [but wrongly, as we will see later!], also a Coriolis field exists, resulting from the components g_{41} , g_{42} , g_{43} of the potential, and being of first order in ω . This field acts orthogonally deflecting on moving masses, and leads to a rotation of the pendulum plane in the Foucault experiment. I have calculated this rotational dragging for the earth [in the Entwurf theory]; it stays far below any observational amount. Such a Coriolis field is also produced by the rotation of the sun and of Jupiter, and it produces secular changes of the orbital elements of the planets and moons, which, however, stay far below the measurement error. On the whole, the perihelion advance of Mercury seems to be the only case where deviations from the classical theory of celestial mechanics are observable today. Nevertheless, the Coriolis fields are nearer to observation than your correction terms to g_{44} because these latter terms have the same symmetry as the correction terms due to oblateness.

Thirring should of course have known of this Coriolis force, either from his courses in classical mechanics or from Einstein’s Vienna speech, where Thirring was present. The first entries in Thirring’s notebook after the receipt of Einstein’s letter deal then exactly with this Coriolis force inside a rotating hollow sphere and outside a rotating full sphere. Later parts of the notebook contain, among other things, drafts of the paper (Thirring 1918) and of sections 1–2 of the paper (Lense and Thirring 1918). From these and other documents (Pfister 2007), it is evident that sections 3–4 of the paper (Lense and Thirring 1918), containing the transformation of Thirring’s results to the orbital elements of the planets and moons of the solar system, are the only contributions by Josef Lense. In detail, the paper (Thirring 1918) contains the metric and the geodesic equation at points near the origin ($r \ll R$) of a rotating spherical mass shell in first order of M/R and up to second order in ω . The calculations are quite tedious and without any elegance. In vector notation (not used by Thirring!), the force on test masses m at position \mathbf{r} , having velocity \mathbf{v} , is in order ω

$$\mathbf{K}_C = -\frac{8mM}{3R}(\boldsymbol{\omega} \times \mathbf{v}). \quad (5.1)$$

In order ω^2 , Thirring finds the force

$$\mathbf{K}_Z = -\frac{mM}{3R}[\boldsymbol{\omega} \times (\boldsymbol{\omega} \times \mathbf{r}) + 2(\boldsymbol{\omega} \cdot \mathbf{r})\boldsymbol{\omega}]. \quad (5.2)$$

The paper (Lense and Thirring 1918) contains the metric and the geodesic equation for the far field ($r \gg R$) of a rotating full sphere in first orders of M/R and ω , resulting in the so-called Lense–Thirring force,

$$\mathbf{K}_{LT} = 2m\mathbf{v} \times \mathbf{H}, \quad \text{with} \quad \mathbf{H} = \frac{2MR^2}{5r^3} \left[\boldsymbol{\omega} - 3 \frac{(\boldsymbol{\omega} \cdot \mathbf{r})\mathbf{r}}{r^2} \right]. \quad (5.3)$$

It is very strange that these papers by Thirring and Lense contain no reference whatsoever to Einstein's decisive letter (Schulmann et al. 1998, Document 369).

5.4 Deficiencies of the 1918 Papers by Thirring and Lense and Their Corrections

The terms of first order in $\boldsymbol{\omega}$ in equations (5.1) and (5.3) are mathematically correct. However, the deficiency here lies in the fact that they are derived only for positions $r \ll R$ and $r \gg R$, respectively. This is a particularly severe deficiency for the so-called Lense–Thirring force (5.3) which, contrary to the usual claims in the literature, is not directly applicable to modern experiments trying to measure this new “gravitomagnetic” force in the field of the rotating earth (more than 90 years after its prediction by Einstein, Thirring, and Lense): The LAGEOS satellites orbit the earth at a height no greater than $r/R \approx 1.92$, and the “Lense–Thirring motion” of 31 milliarcsec/year of their nodes was confirmed with a precision of 10% (Ciufolini and Pavlis 2004), the main problem here being the subtraction of the overwhelming contribution of 120 degrees/year due to the earth's quadrupole. An improvement of accuracy using 3.5 years of laser-ranged observations of the LAGEOS and LARES (launched in 2012) satellites, together with the enhanced Earth gravity field model produced by the space geodesy mission GRACE, yields an (interim) 5% estimate of systematic errors (Ciufolini et al. 2016). The Gravity Probe B satellite (Everitt et al. 2001) was launched in 2004, orbits the earth in a low orbit with $r/R \approx 1.10$, and is destined to measure the precession of 42 milliarcsec/year of gyroscopes inside the satellite. Due to unexpected problems with the electric loading of these gyroscopes, the Lense–Thirring precession has been observed with an accuracy of approximately 19% only (Everitt et al. 2011). Concerning the relevance of equation (5.3) for these experiments, only many decades after Thirring's work was it seen that equation (5.3) is indeed valid not only for $r \gg R$ but for all $r > R$ and equivalently that equation (5.1) is valid for all $r < R$. To my knowledge, the first clear statement of this fact, at least for equation (5.1), appears in Brill and Cohen (1966). Mathematically, this can be seen either by the (somewhat tedious) proof that all higher-order contributions in R/r to equation (5.3) are zero, or, more elegantly, by a symmetry argument: A first-order rotational perturbation of a spherical system can produce only a pure dipole field proportional to r^{-3} .

Coming to the term of order ω^2 in equation (5.2), the paper (Thirring 1918) has to be criticized first in a more general and qualitative way because it puts more weight on these “centrifugal” effects than on the order ω Coriolis effects, notwithstanding Einstein's suggestion to the contrary. And nowhere does it estimate

the order of magnitude of these ω^2 -order terms which are smaller by a factor $(M/R) \cdot (v/c)^2 \approx 10^{-21}$ than Newtonian effects near the rotating earth and therefore far from experimental confirmation even today. Since equation (5.2) is also derived only for $r \ll R$, it has to be mentioned that here the correction terms of higher order in r/R would be nonzero. However, since formula (2) suffers from many other deficiencies, it is not worthwhile calculating such correction terms.

A first technical error in result (5.2) was already found in 1920 by Max von Laue and Wolfgang Pauli and was corrected by Thirring in an erratum to (Thirring 1918): Due to an incorrect volume integration in the derivation of equation (5.2), the prefactor $mM/3R$ has to be changed to $4mM/15R$. A more severe, physical defect of Thirring's model was observed and corrected in the paper (Lanczos 1923): Thirring had modeled the material of the shell as dust, which is wrong because of the centrifugal stresses in the rotating shell, with the consequence that Thirring's solution does not fulfil Einstein's field equations in the shell. The correction of this error leads to a further reduction of the prefactor of equation (5.2) to the value $2mM/15R$.

More than three decades later, Lanczos' arguments were partly repeated in Bass and Pirani (1955) but presented in more mathematical detail and generalized to a spherical shell with latitude-dependent mass density $\rho(\vartheta) = \rho_0(1 + N\omega^2 R^2 \sin^2 \vartheta)$, with constants ρ_0 and N . (At the same time, and obviously independently, similar but less complete results were derived in Hönl and Maue 1956.) Bass and Pirani call the choice $N = -1$ the most interesting case because then the special-relativistic mass increase of the equatorial parts of the rotating shell is exactly compensated, and the whole force \mathbf{K}_Z of equation (5.2) vanishes identically. This result also shows that Thirring's conjecture that only the axial component of \mathbf{K}_Z results from such a special-relativistic mass increase of the equatorial parts of the shell is wrong. [Compare Einstein's letter (Schulmann et al. 1998, Document 369) to Thirring.] In Bass and Pirani (1955), it is also mentioned that in Thirring's model, the self-interaction of the shell (being proportional to M^2) is neglected, but no attempt is made to extend the model beyond the weak-field approximation. Another argument, calling for a treatment of the rotating mass shell at least up to order M^2 , was presented in Soergel-Fabrizius (1960): As already discussed in Thirring (1918), it is possible to eliminate the Coriolis force inside the rotating mass shell by a transformation to a coordinate system counter-rotating with $\tilde{\omega} = -(4M/3R)\omega$. Soergel-Fabrizius now adds the argument that the centrifugal force can vanish in the same reference frame, as it should according to Mach's demand for relativity of rotation, at best if it is of the order $\tilde{\omega}^2 \sim (M^2/R^2)\omega^2$, instead of the order $(M/R)\omega^2$ in Thirring (1918).

A treatment of the rotating mass shell, even exactly in M , was started in Brill and Cohen (1966), by considering a rotational perturbation not of Minkowski space-time but of the Schwarzschild solution. However, they confined themselves to the first order in ω , and derived (for the whole, flat interior of the shell!) a Coriolis-type acceleration, with dragging factor

$$d = \frac{4\alpha(2 - \alpha)}{(1 + \alpha)(3 - \alpha)}, \quad (5.4)$$

with $\alpha = M/2R$, and where R denotes the shell radius in isotropic coordinates. In the weak-field limit $M \ll R$, this dragging factor coincides of course with Thirring's result $d = 4M/3R$. But the central new result of Brill and Cohen (1966) is that in the collapse limit $R \rightarrow M/2$, the dragging factor attains the value $d = 1$. This signifies—within the model class of rotating mass shells and up to first order in ω —a complete realization of the Machian postulate of relativity of rotation: In the collapse limit, the interior of the shell ties off (as a type of separate universe) from the exterior space-time, and interior test bodies and inertial frames are dragged along with the full angular velocity ω of the shell. It has, however, to be admitted that near the collapse limit the shell material attains somewhat unphysical properties: For $R < 3M/4$, the dominant energy condition (Hawking and Ellis 1973) is violated, and in the final collapse limit, the stresses in the shell even diverge.

5.5 The Solution of the Centrifugal Force Problem

All the corrections and extensions of Thirring's work, described in Section 5.4, have left one decisive question open: Is it possible to model a rotating mass shell such that in its whole interior correct Coriolis and centrifugal forces, and no other forces, are induced, hereby fulfilling Mach's postulate of relativity of rotation, in this model class, at least in second order, or in all orders of ω ? The positive answer to this question in Pfister and Braun (1985) rests on two "new" ideas, which could and should have been brought forward already in Thirring's time but which, for inexplicable reasons, were overlooked by all authors before 1985:

- a) Any physically realistic, rotating body will suffer a centrifugal deformation in orders ω^2 and higher and cannot be expected to keep its spherical shape.
- b) If we aim and expect to realize quasi-Newtonian conditions in the interior of the rotating mass shell, with correct Coriolis and centrifugal forces (and no other forces!), the interior of the mass shell obviously has to be a flat piece of space-time. In first order of ω , this flatness is more or less trivial. In contrast, in order ω^2 , this flatness is by no means trivial, and it is indeed violated for Thirring's solution, due to the axial component of his "centrifugal force." Moreover, if Thirring had extended his calculations to orders $\omega^3, \omega^4, \dots$, he would have obtained additional forces in the interior of the rotating mass shell, in conflict with Newtonian physics in a rotating reference system.

With these observations, the problem of a correct centrifugal force inside a rotating mass shell boils down to the question of whether it is possible to connect a rotating flat interior metric through a mass shell (with, to begin with, unknown geometrical and material properties) to the non-flat but asymptotically flat exterior metric of a rotating body. In full generality, this would represent a mathematically

quite intricate free-boundary-value problem for the stationary and axisymmetric Einstein equations. However, if we confine ourselves to a perturbation expansion in the angular velocity ω , all metric functions can be expanded in spherical harmonics, or, due to the axial symmetry, just in Legendre polynomials $P_l(\cos \vartheta)$, where in order ω^n the index l is limited by $l \leq n$. In this way, the Einstein equations reduce to a system of ordinary differential equations for the functions $f_l^{(i)}(r)$ multiplying $P_l(\cos \vartheta)$ ($i = 1, \dots, 4$, for the four different metric coefficients describing the stationary and axisymmetric space-time).

According to Pfister and Braun (1985), to order ω^2 , the shell geometry is given by $r_S = R(1 + \omega^2 c_2 P_2(\cos \vartheta))$, with a constant c_2 and with corresponding corrections in higher (even) orders ω^{2n} . Furthermore, it turns out (Pfister and Braun 1986) that to order ω^3 , the flatness of the interior space-time can only be maintained if the shell material rotates differentially, $\omega_S = \omega(1 + \omega^2 e_2 P_2(\cos \vartheta))$, with a constant e_2 and with corresponding corrections in higher (odd) orders ω^{2n+1} . Surprisingly, the flatness condition enforces a prolate form of the shell: invariant equatorial circumference smaller than the invariant polar circumference. The conditions that the exterior metric (written in the isotropic coordinate r) is asymptotically flat, and joins, at $r = r_S$, continuously to the interior flat metric, lead (for given M and R) to a unique determination of the constants c_{2n} and e_{2n} and of the functions $f_l^{(i)}(r)$ (Pfister 1989). The energy-momentum tensor $T_{\mu\nu}$ of the shell material results then uniquely from the discontinuities of the radial derivatives of $f_l^{(i)}(r)$ at $r = r_S$, and $T_{\mu\nu}$ is ϑ -dependent in orders ω^2 and higher, as already guessed in Bass and Pirani (1955). Only in the collapse limit is the rotating shell with flat interior spherical and rigidly rotating, and it produces the Kerr geometry in the exterior, as was already deduced in de la Cruz and Israel (1968). For a mass shell which deviates from sphericity already in zeroth order of ω , there is no solution with flat interior (Pfister 1989). A summary of these results, using the more elegant, mainly geometric Israel-formalism, and correcting some minor errors, can be found in Pfister and King (2015), Appendix B.

5.6 A Quasiglobal Principle of Equivalence

The success with the proof of “relativity of rotation” inside a rotating mass shell encouraged me to conjecture a generalization of this result in the form of a “quasiglobal principle of equivalence” in general relativity in Pfister and Braun (1985). In short form, this can be stated as “Every acceleration field can—in a finite space-time region—be understood as a gravitational field.” (The reversal is of course not true: A general gravitational field, with curvature, cannot be understood as an acceleration field.) In more detail, and in an operational manner, the conjecture may be formulated in the following way: If a finite laboratory (which we understand as a flat piece of space-time) is arbitrarily accelerated (relative to the “fixed stars”), very general inertial forces (generalizations of Coriolis and centrifugal forces) arise. The

hypothesis then claims that the same forces can be produced in a static laboratory by appropriate and appropriately accelerated masses outside the laboratory, e.g., in a mass shell. Historically, it is interesting that similar ideas were exchanged between Albert Einstein, Paul Ehrenfest, and Gustav Mie already in the years 1912–1913. In a letter to Einstein (Klein et al. 1993, Document 411), Ehrenfest formulates: “Are there nontrivial cases where the equivalence principle is valid also for a finite spatial region? (Macro-equivalence)” and “Determine the most general motions of a laboratory such that the macro-equivalence is valid.” Mie remarks in the discussion of Einstein’s Vienna speech (Einstein 1913):

Imagine that you are driving in a railway coach which is sealed off from the exterior world. You are shaken and jolted in this coach, and you interpret these forces, exerted on your body, as inertial forces resulting from the irregular undulations. The general principle of relativity in Einstein’s formulation would now claim that it is possible to assume a system of gravitating masses which perform accelerated motions around a static railway coach, and thereby exert the same forces on your body as those which you interpret as inertial forces.

But in the years 1912–1913, all these persons were quite skeptical as to whether such a “macro-equivalence” is realized in nature.

Due to the quite general status of our “quasiglobal principle of equivalence,” it will be difficult to give a general mathematical proof. However, recently we succeeded in giving a proof for the case of a linearly accelerated mass shell (Pfister et al. 2005). Since general accelerations can, at least in principle, be combined from linear and rotational accelerations, we regard this result as an important argument for the validity of our principle in general relativity.

A quite general and severe problem with linearly accelerated bodies is that they need—in contrast to rotating bodies—a perpetual supply of energy in order to maintain the acceleration. And since in general relativity the equations of motion of bodies are already contained in the field equations, the energy source (or the motor of the accelerated system) has to be included in the considered system in order to deal with a self-consistent problem. This difficulty may be the reason why, besides the historical paper (Einstein 1912a), only a few articles (see references [18–21] in Pfister et al. 2005) treat (or claim to treat) dragging effects due to linearly accelerated masses. And even these papers compare only quite poorly to the rotating systems of Thirring (1918), Brill and Cohen (1966), and others, because they treat only the weak-field case or contain special relations between mass M and charge q of the shell. Furthermore, in some of these papers, the source of acceleration is not really fixed or is removed to infinity. And in no model considered was the geometry inside the shell guaranteed to be flat, so that pretended dragging effects cannot be clearly distinguished from local gravitational effects due to curvature.

In the model of Pfister et al. (2005), the linear acceleration of the mass shell is accomplished electromagnetically: A spherical Reissner–Nordström shell with mass M , radius R , and charge q is surrounded by a weak, axisymmetric, dipolar, exterior charge distribution $\lambda\sigma(r)\cos\vartheta$, with appropriate fall-off behavior so that the dipole moment is finite and the system is asymptotically flat. The system is chosen time-symmetric around $t = 0$ and is treated only up to terms of order t^2 , and the solutions of the Maxwell equations (on the Reissner–Nordström background) are

chosen such that up to order t^2 no magnetic fields arise. In analogy to the rotating shells in Thirring (1918) and Brill and Cohen (1966), the geometrical and material properties of the shell are chosen such that, in first order of λ , the shell stays rigid and attains no correction terms to its energy density and pressure. It could then be shown in Pfister et al. (2005) that these conditions produce a unique solution of the Einstein–Maxwell equations, with a flat interior of the shell. The linear dragging of electrically neutral test particles inside the shell compares nicely with the results for rotating shells:

- a) In the weak-field limit $M/R \ll 1$, and $q/R \ll 1$, and for the simplest power law charge distribution $\sigma(r) \sim r^{-5}$, having a finite dipole moment, the dragging factor coincides with Thirring’s result $d = 4M/3R$.
- b) For arbitrary values $2M/R \leq 1$, but small q , the dragging factor d has a dependence on M/R similar to that for the rotating mass shell of Brill and Cohen (1966). In particular, in the collapse limit $2M/R \rightarrow 1$, we have $d \rightarrow 1$, i.e., total dragging for arbitrary charge distributions $\sigma(r)$.
- c) In the general strong-field case, the dependence of d on M/R and q/R compares reasonably with the results for a rotating, charged mass shell in Pfister and King (2002). In particular, also here d can attain negative values (antidragging) in part of the parameter region where energy conditions (for the shell material) are violated.

5.7 Cosmological Considerations

The solution of the centrifugal force problem in Section 5.5 and herewith the complete realization of Mach’s postulate of relativity of rotation in the model class of rotating mass shells, is still unsatisfying for at least one reason: As the quotation from Mach (1872) in Section 1 makes clear, Mach voted for a realization of his postulate in our real universe, for which the models of mass shells with their asymptotically flat boundary conditions are only insufficient substitutes.

A first successful extension of the work of Thirring (1918) and Brill and Cohen (1966) to cosmological boundary conditions was performed in Lewis (1980) and in more detail in Klein (1993). There it was shown that a rotating mass shell with flat interior can also be embedded into a rotationally disturbed FLRW cosmology. The dragging effects inside the mass shell are similar to the results for isolated shells but depend of course also on the type of cosmology (open or closed) and on the cosmic mass density. More recently, analyses of rotational perturbations of pure FLRW cosmologies (without the somewhat unrealistic mass shells) have appeared in Bičák et al. (2004) and Schmid (2006, 2009), again with a confirmation of the typical Machian dragging effects.

In the cosmological (and therefore time-dependent) context, it is also important to state that the Machian effects in general relativity are not causal but are produced by the time-independent constraints of the Einstein equations, as was already

conjectured in Wheeler (1964) and finally proven in Lynden-Bell et al. (1995). Recently, the instantaneous effect of dragging not only due to accelerated matter but also by (rotating) gravitational waves on local inertial frames has been shown by Bičák et al. (2012). Within a first-order rotational perturbation, this fact is even more or less trivial because such perturbations have dipolar characteristics, whereas causal signals in general relativity would have quadrupolar characteristics. A somewhat provocative but very illuminating manifestation of such non-retarded Machian effects, even in daily life, can be expressed in a bonmot in Schücking (1996):

Mach's principles—whatever they may be—will always find their defenders and believers. When one of its promoters, Dennis Sciama, slammed on the brakes of his car, propelling his girlfriend, seated next to him, toward the windshield, she was to be heard moaning, “All those distant galaxies!”

Finally, I should like to summarize the present observational confirmation of the main theoretical topic of this article, “relativity of rotation,” in our universe. Here I have in mind the following type of experiment: A physicist in a closed laboratory determines the inertial axes (relative to the laboratory walls) by all possible methods, and as precisely as possible. Then he opens the “windows” of the laboratory, looks to the sky, and is (hopefully) surprised beyond all expectation that these local inertial axes are non-rotating relative to the distant galaxies, quasars, and the cosmic background radiation. (In the textbook (Misner et al. 1973), this fact is called a “miracle of miracles,” and in the textbook (Weinberg 1972), it is named “surely a remarkable coincidence.”) What is then the present accuracy of this “cosmic coincidence” for different sizes of the “laboratory”? Concerning real laboratory-size instruments, the most precise ones seem to be laser gyroscopes, and with these the coincidence is tested with an accuracy of 10^{-8} of the earth's angular velocity ω_E (Stedman 1997). For terrestrial reference systems, realized, e.g., by VLBI and GPS, the accuracy is $10^{-9}\omega_E$ (Kovalevsky et al. 1989). For the dynamical solar reference system, in which the orbits of the planets optimally obey the relativistically corrected Newton laws, the accuracy is $5 \cdot 10^{-9}\omega_E$ (Kovalevsky et al. 1989). For the galactic reference frame, realized by the Hipparchos catalogue, it amounts to $7 \cdot 10^{-8}\omega_E$ (Kovalevsky et al. 1997), which can possibly be improved by a factor of 200 through the planned GAIA mission. The number which does not directly measure the relative rotation between reference systems, but which nevertheless is of interest here, is the vorticity of the cosmic background radiation which is limited by $10^{-21}\omega_E$ (Kogut et al. 1997), amounting to less than 10^{-8} “revolutions” in the age of the universe.

Obviously, these data represent a remarkable “influence” of the cosmos on our local physics. The only other such connection, being seriously discussed today, takes place between the local “time arrow” and a low-entropy state of the early universe.

References

- Bass, L., & Pirani, F. A. E. (1955). On the gravitational effects of distant rotating masses. *Philosophical Magazine*, *46*, 850–856.
- Bičák, J., Katz, J., Ledvinka, T., & Lynden-Bell, D. (2012). Effects of rotating gravitational waves. *Physical Review*, *D85*, 124003.
- Bičák, J., Lynden-Bell, D., & Katz, J. (2004). Do rotations beyond the cosmological horizon affect the local inertial frame? *Physical Review*, *D69*, 064011.
- Brans, C. (1962). Mach's principle and a relativistic theory of gravitation. II. *Physical Review*, *125*, 2194–2201.
- Brill, D. R., & Cohen, J. M. (1966). Rotating masses and their effect on inertial frames. *Physical Review*, *143*, 1011–1015.
- Ciufolini, I., & Pavlis, E. (2004). A confirmation of the general relativistic prediction of the Lense–Thirring effect. *Nature*, *431*, 958–960.
- Ciufolini, I., Paolozzi, A., Pavlis, E. C., Koenig, R., Ries, J., Gurzadyan, V., et al. (2016). A test of general relativity using the LARES and LAGEOS satellites and a GRACE Earth gravity model: Measurement of Earth's dragging of inertial frames. *The European Physical Journal*, *C76*, 120.
- de la Cruz, V., & Israel, W. (1968). Spinning shell as a source of the Kerr metric. *Physical Review*, *170*, 1187–1192.
- Einstein, A. (1912a). Lichtgeschwindigkeit und Statik des Gravitationsfeldes. *Annalen der Physik*, *38*, 355–369.
- Einstein, A. (1912b). Zur Theorie des statischen Gravitationsfeldes. *Annalen der Physik*, *38*, 443–458.
- Einstein, A. (1912c). Gibt es eine Gravitationswirkung, die der elektrodynamischen Induktionswirkung analog ist? *Vierteljahrsschrift für gerichtliche Medizin und öffentliches Sanitätswesen*, *44*, 37–40.
- Einstein, A. (1913). Zum gegenwärtigen Stande des Gravitationsproblems. *Physikalische Zeitschrift*, *14*, 1249–1266.
- Einstein, A., & Grossmann, M. (1913). *Entwurf einer verallgemeinerten Relativitätstheorie und einer Theorie der Gravitation*. Leipzig and Berlin: Teubner.
- Everitt, C. W. F., Buchman, S., DeBra D. B., Keiser, G. M., Lockhart, J. M., Muhlfelder, B., et al. (2001). Gravity Probe B: Countdown to launch. In C. Lämmerzahl, C. W. F. Everitt & F. W. Hehl (Eds.), *Gyros, clocks, interferometers . . . : Testing relativistic gravity in space*. Lecture Notes in Physics (Vol. 562, pp. 52–82). Berlin: Springer.
- Everitt, C. W. F., DeBra, D. B., Parkinson, B. W., Turneare, J. P., Conklin, J. W., Heifetz, M. I., et al. (2011). Gravity Probe B: Final results of a space experiment to test general relativity. *Physical Review Letters*, *106*, 221101.
- Hawking, S. W., & Ellis, G. F. R. (1973). *The large scale structure of space-time* (p. 91). Cambridge: Cambridge University Press.
- Hönl, H., & Maue, A. W. (1956). Über das Gravitationsfeld rotierender Massen. *Zeitschrift für Physik*, *144*, 152–167.
- Klein, C. (1993). Rotational perturbations and frame dragging in a Friedmann universe. *Classical and Quantum Gravity*, *10*, 1619–1631.
- Klein, M. J., Kox, A. J., & Schulmann, R. (1993). *The collected papers of Albert Einstein* (Vol. 5). Princeton: Princeton University Press.
- Klein, K. J., Kox, A. J., Renn, J., & Schulmann, R. (1995). *The collected papers of Albert Einstein* (Vol. 4). Princeton: Princeton University Press.
- Kogut, A., Hinshaw, G., & Banday, A. J. (1997). Limits to a global rotation and shear from the COBE DMR four-year sky maps. *Physical Review*, *D55*, 1901–1905.
- Kovalevsky, J., Mueller, I. I., & Kolaczek, B., (Eds.). (1989). *Reference frames in astronomy and geophysics*. Dordrecht: Kluwer Academic.

- Kovalevsky, J., Lindegren, L., Perryman, M. A. C., Hemenway, P. D., Johnston, K. J., Kislyuk, V. S., et al. (1997). The Hipparcos catalogue as a realization of the extragalactic reference frame. *Astronomy and Astrophysics*, 323, 620–633.
- Lanzos, K. (1923). Zum Rotationsproblem der allgemeinen Relativitätstheorie. *Zeitschrift für Physik*, 14, 204–219.
- Lense, J., & Thirring, H. (1918). Über den Einfluss der Eigenrotation der Zentralkörper auf die Bewegung der Planeten und Monde nach der Einsteinschen Gravitationstheorie. *Physikalische Zeitschrift*, 19, 156–163. [English translation in *General Relativity and Gravitation*, 16 (1984): 727–741].
- Lewis, S. M. (1980). Machian effects in nonasymptotically flat space-times. *General Relativity and Gravitation*, 12, 917–924.
- Lynden-Bell, D., Katz, J., & Bičák, J. (1995). Mach's principle from the relativistic constraint equations. *Monthly Notices of the Royal Astronomical Society*, 272, 150–160.
- Mach, E. (1872). *Die Geschichte und die Wurzel des Satzes von der Erhaltung der Arbeit* (p. 48). Prag: Calve.
- Misner, C. W., Thorne, K. S., & Wheeler, J. A. (1973). *Gravitation* (p. 547). San Francisco: Freeman.
- Pfister, H. (1989). Rotating mass shells with flat interiors. *Classical and Quantum Gravity*, 6, 487–503.
- Pfister, H. (2007). On the history of the so-called Lense–Thirring effect. *General Relativity and Gravitation*, 39, 1735–1748.
- Pfister, H., & Braun, K. H. (1985). Induction of correct centrifugal force in a rotating mass shell. *Classical and Quantum Gravity*, 2, 909–918.
- Pfister, H., & Braun, K. H. (1986). A mass shell with flat interior cannot rotate rigidly. *Classical and Quantum Gravity*, 3, 335–345.
- Pfister, H., Frauendiener, J., & Hengge, S. (2005). A model for linear dragging. *Classical and Quantum Gravity*, 22, 4743–4761.
- Pfister, H., & King, M. (2002). Rotating charged mass shell: Dragging, ant dragging, and the gyromagnetic ratio. *Physical Review*, D65, 084033.
- Pfister, H., & King, M. (2015). *Inertia and gravitation: The fundamental nature and structure of space-time*. Lecture Notes in Physics (Vol. 897). Heidelberg: Springer.
- Schmid, C. (2006). Cosmological gravitomagnetism and Mach's principle. *Physical Review*, D74, 044031.
- Schmid, C. (2009). Mach's principle: Exact frame-dragging via gravitomagnetism in perturbed Friedmann–Robertson–Walker universes with $K = \pm 1, 0$. *Physical Review*, D79, 064007.
- Schücking, E. L. (June, 1996). Gravitation and inertia. *Physics Today*, 6, 58.
- Schulmann, R., Kox, A. J., Janssen, M., & Illy, J. (1998). *The collected papers of Albert Einstein* (Vol. 8). Princeton: Princeton University Press.
- Soergel-Fabricsius, C. (1960). Thirring-Effekt im Einsteinkosmos. *Zeitschrift für Physik*, 159, 541–553.
- Stachel, J. (1980). Einstein and the rigidly rotating disk. In A. Held (Ed.), *General relativity and gravitation* (Vol. 1, pp. 48–62). New York: Plenum Press.
- Stedman, G. E. (1997). Ring-laser tests of fundamental physics and geophysics. *Reports on Progress in Physics*, 60, 615–688.
- Thirring, H. (1917). Wirkung rotierender Massen. Notebook. Österr. Zentralbibliothek, Universität Wien.
- Thirring, H. (1918). Über die Wirkung rotierender ferner Massen in der Einsteinschen Gravitationstheorie. *Physikalische Zeitschrift*, 19, 33–39; erratum in *Physikalische Zeitschrift* 22 (1921): 29–30 [English translation in *General Relativity and Gravitation*, 16 (1984): 712–727].
- Thirring, H. (1966). Ernst Mach als Physiker. *Almanach der Österreichischen Akademie der Wissenschaften*, 116, 361–372.
- Weinberg, S. (1972). *Gravitation and cosmology* (p. 17). New York: Wiley.
- Wheeler, J. A. (1964). Mach's principle as boundary condition for Einstein's equations. In H. Y. Chiu & W. F. Hoffmann (Eds.), *Gravitation and relativity* (pp. 303–349). New York: Benjamin.

Chapter 6

Relativistic Lighthouses: The Role of the Binary Pulsar in Proving the Existence of Gravitational Waves



Daniel Kennefick

6.1 Introduction

In 1993 Joseph Taylor and Russell Hulse received the Nobel Prize for their discovery of the first binary pulsar, PSR 1913+16. Their citation acknowledged the importance of their work for the field of gravitation, and the accompanying press release stressed the special significance of their measurements of the orbital motion of the system in providing the first experimental evidence for the existence of gravitational waves (RSAS 1993). Nobel Prizes for work in astronomy and astrophysics were once very rare, and prizes awarded for work in gravitational physics have been even rarer. Einstein himself was denied any citation for his discovery of general relativity. This Nobel Prize is therefore a striking demonstration of the importance of gravitational waves as a topic in physics. It is all the more interesting then that gravitational wave research had been, until Taylor and Hulse's discovery, a marginal and controversial field. The very existence of gravitational waves, even as a theoretical prediction of Einstein's theory, had frequently been doubted in the half century before Taylor and Hulse's work. The story of how this classic experiment by two astrophysicists settled a long-standing theoretical controversy seems a natural for study from a historical

Parts of this paper appeared previously in an article in the *European Physical Journal H* entitled "The binary pulsar and the quadrupole formula controversy" Kennefick, D. EPJ H (2017). 42, 293–310.

D. Kennefick (✉)
Faculty of Physics, University of Arkansas, Fayetteville, AR, USA
e-mail: danielk@uark.edu

perspective. It is all the more interesting at a time when a second Nobel Prize for work on gravitational waves is widely anticipated, following the remarkable first signal detected by the LIGO/VIRGO collaboration, announced in early 2016.

My own interest in the binary pulsar experiment derives primarily from its relation to the long-running quadrupole formula controversy. This centered on the validity, within general relativity, of the quadrupole formula, first derived by Einstein in 1918, when applied to binary star systems such as the binary pulsar (Kennefick 2007). This aspect of the binary pulsar's history, while interesting in itself, is highly relevant to the question of proving the existence of gravitational waves. What Taylor and Hulse achieved was to show that the rate of decay of the binary pulsar's orbit was in agreement with the prediction of the quadrupole formula, suggesting emission of gravitational radiation by the system as the cause of the decay (Hulse and Taylor 1975). Obviously this line of logic depended on the assumption that the quadrupole formula was correctly derived from the established theory.¹ This comes even more into focus when one realizes that the theoretical controversy over gravitational waves, in its early phases, focused on the question of whether gravitational waves could exist at all, or could be emitted by binary stars.

This topic bears somewhat on the long-standing debate between Harry Collins and Allan Franklin over Collins' concept of the experimenters' regress, as this applied to the Weber controversy, coincidentally enough a controversy over the detection of gravitational waves by Earth-based detectors. It is well known that the exchange between Collins and Franklin over gravitational wave detection was seen as a significant episode in the so-called Science Wars. At that time philosophers, like Franklin, debated with sociologists, like Collins, over issues such as realism versus relativism, the demarcation problem in science studies, and so on. The debate between Collins and Franklin centered on whether physicists had rational or objective grounds for closing debates, or whether achieving closure in scientific controversies depended on the social relations between the participants and their community. Sociologists of science like to argue that "interpretive flexibility" means

¹An interesting side note to the binary pulsar discovery of gravitational waves is the common use of phrases like "indirect detection" of gravity waves, to distinguish Taylor and company's work from the long awaited "direct detection" of gravitational waves by Earth-based detectors. A number of people have pointed out the fallacy in this kind of thinking, observing that, strictly speaking, the binary pulsar evidence is no more indirect than any other detection. Two people I am thinking of particularly are Thibault Damour and Allan Franklin, both of whom have made the point to me personally. While it is true that the astronomers use electromagnetic signals from the source system, and must then infer the presence of gravitational waves from the observed behavior, the same is true of Earth-based detectors, which also use electromagnetically controlled detection of local masses and deduce the presence of gravitational waves from the motions of those masses. In some sense the only difference is that the binary pulsar astronomers only observe the source of the gravitational waves, and thus cannot comment on the propagation through space of these waves. Another distinction, of some relevance to our discussion, is that the theory required to analyze the binary pulsar system is not the linearized gravity which suffices for the Earth-based detector, and therefore, it could be argued, it is a more complex and more controversial process of deduction. Certainly the response to the first detection of gravitational waves by the LIGO instruments has been notable for its lack of any skeptical voice.

that physicists always have the option to keep a debate open, but that social cohesion depends upon the ability of the core group to eventually achieve a consensus, even in defiance of the wishes of the remaining group of outsiders who regard the matter as unsettled (Collins 1994, 2004). Franklin's take was that the rump group of outsiders in the Weber controversy were behaving irrationally in refusing to accept comprehensive experimental evidence contradicting their view and had, in some sense, placed themselves outside of the sphere of rational scientific discourse (Franklin 1994).

Inspired by Collins' work, I made use of the analogous concept of the Theoreticians' Regress to explain the intractability of the controversy on the theoretical side of the gravitational wave field. Somewhat to my relief, my own study did not appear to be so controversial, in the context of the science wars, no doubt largely because of my insignificant status in the field. Additionally, neither side claimed that theorists were directly confronted with the objective reality of the laboratory. But what about the fact that the close of the debate in my story was apparently connected, certainly timed so as to suggest a connection, with the arrival of experimental evidence? Was it not particularly satisfactory for "realists" that unambiguous, and largely unchallenged, experimental evidence should help to close out debate among theorists? While I was happy that my study was not likely to play a role in the science wars, I was shy of this one issue that was apparently relevant to the questions at issue in that struggle.

Both Collins and Franklin were pioneers in the careful micro-study of experimental method and practice. Through personal contact with both men, and others, I was inspired to adopt this kind of approach. From my perspective, Collins and Franklin appeared to be saying very similar things. I confess to being somewhat uncomfortable in addressing the precise role played by the binary pulsar in settling the quadrupole formula controversy, lest I be seen to be firing a shot in a war which, however interesting the individual debates, I find slightly incomprehensible. Surely what mattered was that both Collins and Franklin believed in close detailed studies of what scientists actually did, not in whether they agreed on points of principal. Here, I suppose, my own outlook as a historian differed from either Collins or Franklin, who as a sociologist and a philosopher, respectively, viewed the micro-study as a means to an end. Their goal lies not only in the interest of the study itself but also in what it teaches us about the way science is done. At any rate, it seemed as if I was a conscientious objector in the science wars. Now that a ceasefire among those who value the scientific endeavor seems likely to endure (Collins 2009), this paper is by way of being a belated commentary on this issue, from the perspective of a draft dodger now returned to the scene of the fray in peacetime.

6.2 Controversy

The background to the story can be sketched relatively briefly (for a fuller account, see Kennefick (2007)). The theory of gravitational waves dates to 1916 with Einstein's first paper on the subject (Einstein 1916), only half a year after his

publication of the final form of his general relativity theory. In 1918 Einstein published a paper correcting a certain error from the paper of 1916 and presenting, for the first time, the quadrupole formula, expressing the rate of emission of gravitational wave energy by a system of accelerating masses (Einstein 1918).

When Einstein derived the quadrupole formula, it was on the basis of the linearized approximation of general relativity. This permitted him to make the calculation relatively straightforward, because in the coordinate system adopted by him, the linearized equations of gravity take on a form which is directly analogous to the Maxwell equations for electromagnetism, a theory in which the role of radiation was, and is, reasonably well understood. But, since general relativity is a nonlinear theory, this linearized approximation can hold only for very weak fields, which specifically excludes systems, such as a binary star system, which are held together by their own gravitational interaction. Since it is only this type of system which (as far as we know today) might be capable of producing detectable gravitational waves, this approximation leaves something to be desired as far as sources go. It is thought to be ideal for the study of gravitational wave detectors however. The question then is, does the quadrupole formula give a reasonable approximation of the source strength of possible astrophysical sources of gravitational waves, especially binary stars?

Famously, Einstein himself came to entertain doubts about the existence of gravitational waves (indeed, there is evidence that his paper of 1916 was preceded by a brief period of skepticism on the subject; see Kennefick (2007, pp. 44–49)), when he and his then assistant Nathan Rosen came to look for an exact solution of the Einstein equations representing plane gravitational waves (Einstein and Rosen 1937). They discovered that it was not possible to construct a metric in a given coordinate system which did not include a singularity somewhere in the spacetime representing the plane gravitational waves. Subsequently it was shown that this singularity is merely a coordinate singularity, rather than a physical singularity, but at the time, Einstein and Rosen interpreted it as physical, arguing that such spacetimes could not exist. However, before the paper was published, Einstein realized that his argument was mistaken. Nevertheless, in the published version, he still included a discussion of the possibility that binary stars would not emit gravitational waves, in spite of the fact that the quadrupole formula suggests that they would. Einstein's assistant who succeeded Rosen, Leopold Infeld, afterward always insisted that this was Einstein's final word on the subject which, in a strictly published sense, it was.

When interest in general relativity began to pick up again in the mid-1950s, Rosen and Infeld advanced a number of arguments whose common point was that binary star systems would not undergo orbital decay as a result of emitting gravitational waves. Hermann Bondi also entertained serious doubts on this score, arguing that the analogy with electromagnetism which lay behind the original notion of gravitational waves actually pointed this way. His view was that in electrodynamics it was believed that accelerating charges emitted radiation and that the same was expected to hold true in the case of gravity. But since the theory was a theory of general relativity, how did one define what was accelerating? In Bondi's

view, an inertial particle in general relativity was one which followed a geodesic. An accelerating particle was one which did not. Since binary stars in orbit around each other followed the geodesics of the local spacetime, they were not accelerating, in this sense. As particles in a form of inertial motion, their motion would not be of the type which should decay in response to radiation reaction.

These kinds of arguments came up for discussion at a seminal 1957 meeting at Chapel Hill, North Carolina, which was the inspiration for the General Relativity and Gravitation series of meeting which have continued to the present day as the leading conferences in the field of general relativity. The meeting is important for the history of gravitational waves because it was there, in response to arguments raised by Rosen, that Richard Feynman and Bondi himself, responding to the work of Felix Pirani, put forward the “sticky bead” argument that gravitational waves must carry energy. As a result of this, the debate shifted to the question of whether binary stars could emit gravitational waves. This question was still being debated at the third General Relativity and Gravitation meeting held in Warsaw in 1962. Feynman attended this meeting and, one may speculate, was perturbed to find that the questions he had thought were settled in 1957 were still being aired. While the questions had changed somewhat, nevertheless Feynman had, in 1957, made an impassioned case for the field to abandon a “too rigorous” approach as being infertile in theoretical physics (De Witt 1957). In a celebrated letter home from the conference to his wife, Feynman painted a Felliniesque portrait of a physicist trapped inside a field full of “dopes” (126 of them at the conference, according to his letter) rehearsing the same arguments over and over again like “a lot of worms trying to get out of a bottle by crawling all over each other” (Feynman and Leighton 1988).

Ironically enough, it was the work of Bondi himself, as much as of any other relativist, which did the most to convince most relativists that binary stars did indeed decay in their orbits as a result of gravitational wave emission. But the debate seemed of little practical relevance, since the one thing that everyone involved agreed upon was that the rate at which this decay took place was too small for it to be observable in any known orbital system. Very likely it was for this reason that the debate became very quiet in the decade between 1965 and 1975. The discovery of the binary pulsar in late 1974 undoubtedly did much to reinvigorate this debate, which by then had shifted to a new question, whether the quadrupole formula was the correct formula for strong gravity binaries of this kind. Over the course of the following decade, the debate was fairly vigorous, until it petered out in the mid-1980s, when the remaining skeptics grew quiet (again, for a discussion of all of this history, with references, consult Kennefick (2007)).

What is interesting about the role of the binary pulsar in this story is that there are good grounds for believing that its primary role was to stimulate the controversy into new life. It is usually thought of as the agency by which the controversy was settled (and this is certainly a role which is of interest to this paper), but another possible reading is that it actually made the controversy more prominent and more contentious and that this served, with time, to bring it to a conclusion by focusing the attention of theorists upon it. One might speculate that we are dealing

with a controversy downsizing principle, in analogy with the problem of cosmic downsizing in extragalactic astronomy, which revolves around the observation that over time quasars come to have smaller and smaller black holes. Since black holes should only ever grow in size, it is claimed that this observational effect arises because the big ones have already used up all their fuel and “turned off.” The situation is thought to be similar to that which obtains for stars, where the larger stars, which paradoxically contain more fuel, burn the fuel at a far faster rate and live a much shorter life than do less massive stars.

In the case of scientific controversies, we may similarly expect, at any given moment to find many more small and almost moribund controversies than strident ones, because the former will be more long-lived. The fuel which is only slowly consumed in a small controversy is not the number of issues to be debated. I agree with those who think such points are all but inexhaustible. The fuel is the number of potential participants in the controversy. Where the number of participants is low, each of them may feel comfortable conceding a long period of debate to what is a manageable number of colleagues. As the number involved in the controversy rises, the ability to mediate the controversy by direct personal relations between all participants is strained. The consequences of remaining on the fence become less predictable as they become potentially more serious, since more people involved means potentially more influential people having a vested interest in the outcome. The participants come under pressure to take a definitive position and tend to do so more quickly. To continue with the analogy, the fuel is more quickly processed through the various stages, from open-minded participant, to committed protagonist, to close-minded ideologue, at the end of which no further debate is possible. In essence, the controversy which burns most brightly extinguishes itself most quickly. To be sure, I am here merely taking a long-established piece of folk wisdom and dressing it up in academic clothes. The phrase “slow-burning controversy” already nicely encapsulates the image I am trying to convey.

So let us examine briefly the course of the quadrupole formula controversy in the 1970s. We have already summarized the debate over whether binary stars could emit gravitational waves, a debate which flourished in the late 1950s and early 1960s. There then followed a period in which it was regarded as settled, by a large majority, that binary stars did undergo radiation damping as a result of gravitational wave emission. The detail of how this occurred was perhaps not regarded as a terribly pressing problem, given that no one was familiar with any known astronomical systems which, according to the quadrupole formula itself, would undergo a measurable decay in their orbits. The state of affairs bore a close approximation to the situation in controversies which have passed the point of crystallization, which is to say that even though there remained some who doubted the consensus opinion that the quadrupole formula was approximately correct, their views did not receive much public airing. In fact, however, it was still possible for their views to be aired, the problem was simply not important enough for huge notice to be taken of anyone’s views on the matter.

A good example of the status of the debate on the eve of the discovery of the binary pulsar is the June, 1973, Paris meeting on gravitational waves at which

Havas gave a talk outlining his view that the question whether binary stars did emit gravitational waves at all was still unsettled and advancing his critique of the main calculations which agreed with the quadrupole formula result (Havas 1973). In the conference proceedings, two of the remarks in response to Havas' talk can be regarded as sharing his skepticism, two as disagreeing with it, and two as neutral (at least phrased in a neutral way). This certainly suggests not only that Havas had leave to raise such issues with his peers but also that he had an audience part of which, at least, was sympathetic. At the same time, the problem was not at the forefront of theoretical concerns at that moment. It was not considered irrelevant or uninteresting, after all the very fact of the conference being held at all suggests otherwise, but the fact that no astrophysical applications had been discovered certainly restricted its urgency.

Within little over a year, the situation was transformed completely.

6.3 Discovery

Pulsars were discovered in 1967 by Jocelyn Bell and Tony Hewish using the Interplanetary Scintillation Array at the Mullard Radio Astronomy Observatory near Cambridge, England. It quickly became apparent that pulsars were a real-life instance of a long-standing theoretical entity, the neutron star, which had been first proposed by Walter Baade and Fritz Zwicky decades previously, in Baade and Zwicky (1933) (see Haensel et al. (2007, pp. 2–4) for a brief history). The problem of gravitationally collapsed objects becomes of greater theoretical interest following the discovery of quasars by radio astronomers in the 1950s and was further stimulated by the pulsar discovery. By the early 1970s, only a few dozen pulsars were known, and Joe Taylor of the University of Massachusetts, together with his graduate student Russell Hulse, proposed to do a computerized search for them with the large Arecibo dish in Puerto Rico to provide a much larger ensemble of discovered objects. It was a specific aim of Taylor's proposal that such a large number of pulsars might feature one which was part of a binary system (Hulse 1997). This would permit the measurement of the mass of the pulsar, a topic of immense astrophysical interest, since the very idea of neutron stars had arisen following the work of Subramanian Chandrasekhar on the limiting mass of white dwarf stars. That a close binary neutron star system had been suggested as a possible source of detectable gravitational waves as early as 1963 by Freeman Dyson was almost certainly not on Taylor's mind as he began his pulsar search (Dyson 1963). This was all the more true since Dyson's suggestion had been made in the context of an argument that arbitrarily advanced alien civilizations might construct such systems for the purpose of interstellar navigation.

In early July 1974, Hulse, down at Arecibo, recorded a pulsar, just barely strong enough to be detected by the system, unusually sensitive for its day as it was, whose position on the sky automatically baptized it with the name PSR 1913+16. After confirmation that this was indeed a pulsar, including measuring its period, Hulse

recorded the word “fantastic” on his observing record, referring to the fact that the pulsar had the second shortest period known at that time. At this point he had no notion that it was in a binary system, only the rotational period of the neutron star itself had been measured, not its orbital period. The only foretaste of what was to come was that subsequent attempts to confirm that rapid pulse in these first observations did not agree, to Hulse’s frustration. He even went so far as to cross out and erase these subsequent attempts from his log (Hulse 1997).

In late August Hulse returned to this object, in a routine way, to try to confirm its period. As before he found that its period kept changing with each measurement. Indeed, by a curious coincidence, he found that he almost repeated the same set of measurements each time the pulsar came overhead at Arecibo (the dish at Arecibo is so large it is built into a small valley and thus cannot be observed very far from the zenith of the sky). This would turn out to be due to the fact that the pulsar binary has an orbital period of just under 8 hours and thus completes a little over 3 orbits with every rotation of the Earth. It did not take Hulse long to convince himself that he had discovered a pulsar in a binary system, and it was immediately clear to him and to his advisor Taylor that they were dealing with an extraordinary system. An 8-hour orbital period represented an orbiting system involving massive objects with an unprecedentedly small physical separation from each other. Indeed, word got around quickly about the new discovery, to the extent that the first theoretical paper commenting on the binary pulsar appeared in late 1974 (Damour and Ruffini 1974), while the discovery paper itself appeared only the next year in Hulse and Taylor (1975).

There can be little doubt that interest in the radiation problem from binary stars was reinvigorated by the binary pulsar discovery. Here was a real-world example of a system where radiation damping might actually be measurable. Of course there were doubts expressed, on the theoretical side (Damour and Ruffini 1974), that the effect really would be measurable, but the experimenters were nevertheless not ruling it out. In an interview Joe Taylor recalls his own view at the time (interview conducted by the author by phone on 2nd May, 2008):

The person who put us onto that was Bob Wagoner. It happened that once the news was out and it became public that this thing was there and that we were observing it, I responded to a number of invitations to go and give talks about it and ended up making a grand tour around North America where I made five or six stops and one of them was at Stanford and Bob Wagoner there actually gave me his paper predicting the orbital period decay to carry back with me since he knew I was going to be at Harvard a couple of days later and I handed it to Alex Dalgarno the editor of ApJ Letters. So it was Bob’s paper (Wagoner 1975) that I first began to take seriously and to recognize that with the current state of the art then, in October 1974 of doing pulsar timing, it was clear that, if his numbers were right, and I assumed they were, it would take us a number of years to see any effect, but not an unreasonable number and if we could improve the timing accuracy a little bit it might happen even sooner and that’s more or less what happened.

While relativists were excited about a number of tests of general relativity which could be made for this system whose components were moving under the influence of unprecedentedly strong gravitational forces, it seems that the measurement of the

binary pulsar orbital decay came significantly earlier than most people expected, as Taylor agrees (interview, 2nd May, 2008):

I think that's right and that's largely because at that time it wasn't yet recognized that doing really high precision timing of pulsar signals was a very important goal.

Nevertheless the possibility was in the air from late 1974 onward, and the fact that it would take a significant amount of time gave the theorists ample time in which to apply new techniques and increased effort to the problem of analyzing the orbital evolution of such a system as it responded to its own gravitational wave emission.

To what extent was this activity on the theoretical side visible to the experimenters? Given that their result, when available, was likely to have a decisive effect on the controversy, it is remarkable that they went totally unaware of it until they finally had a result to announce. This announcement was made, in its earliest version, at the ninth Texas Symposium on Relativistic Astrophysics in Munich in 1978. The Texas series of meetings had a tradition of announcements of important observational results. The first Texas meeting had been held in response to the growing interest in quasars as new objects discovered by radio astronomers in the late 1950s. Taylor's talk in Munich is one of the more celebrated of the announcements made at this series of meetings (interview, 2nd May, 2008).

Well, I'll tell you when I first even knew that there was any debate, was at the Texas Symposium in Munich.² And so somebody asked me a question, well let me back up just a little bit. I was scheduled to give a paper there on something like the second or third day of the conference, and Jürgen Ehlers, who was one of the conference organizers, recognized that somehow not getting to this until nearly the last day of the conference was not a good idea. So he asked me to get up and say just a few words about it in a session on the first day so that at least people would know what I looked like and we could talk in the halls, and so forth, afterwards. So I did that and I basically gave the result and said I'll give all the details at the scheduled time the day after tomorrow, or something like that. Somebody then in the audience asked a question, I don't remember who it was, 'when you say that you have seen the period decay and it agrees with the prediction, what prediction are you using?' And I sort of was blind-sided by that. I just thought that everyone knew how to calculate this, except maybe me. And so I think I must have stood there wondering how to answer for a minute and Tommy Gold, who happened to be the session chairman, whispered in my ear, 'Landau and Lifshitz,' so I said it's given in Landau and Lifshitz. So that more or less is what transpired. I mean, I remember having conversations later with people about it and I began to realize that, of course, that was just sort of an heuristic formula and the calculation wasn't even derived, I guess, in Landau and Lifshitz, it was given as an exercise for the student to do.

It is humorous to note that Gold, the session chairman, had been, with his collaborator Bondi, one of the early skeptics concerning whether binary stars could emit gravitational radiation. Although Gold would certainly have been very familiar with Landau and Lifshitz's treatment, he might also have been inclined to agree with Bondi's comment that it was very "glib."

²At this point on the interview recording, the author can hear himself say 'Really.'

So once Taylor was apprised of the existence of the controversy, what was his reaction? (interview, 2nd May, 2008):

So ok, so I was aware then that there was a controversy about it. Whenever I quizzed theorists, that I knew pretty well, about it, they tended to be people like Kip Thorne, for example. Kip always said, 'oh yes, you know, we're still worrying about the mathematical details, but we know its right.' And my impression was that, I think pretty much I gained the impression that you convey to a large extent in your book as well,³ that the more mathematically oriented physicists, and particularly those who had been doing relativity in mathematics departments, were still concerned about the lack of rigor and the full mathematical beauty, but the physicists like Thorne and Feynman and others just had little patience with that kind of concern and wanted to get on with it and see what you could do with it. And they more or less told me 'don't worry about it.'

So communication between theorists and experimenters contained this interesting feature that a reasonably lively controversy among the theorists could be completely invisible to the experimenters. Obviously the controversy was not one which consumed the total energy of theorists in the field, but it still involved a good deal of back and forth and even a dedicated workshop, during the period in question, and yet no mention was made of its existence within Taylor's hearing. Partly, as Taylor says, this was because of the kind of theorists he was talking to. In the field of relativistic astrophysics, there were people close to the astrophysics end of the spectrum, and people closer to the relativity end, and Taylor, as an astrophysicist, was naturally more likely to talk to those on the astrophysics end. Since those theorists were less likely to be skeptical of the quadrupole formula, they naturally chose not to bring up any caveats about the derivations which they felt were unlikely ever to have a bearing on the observations underway. Furthermore, and this bears on a point I will try to bring out at the end of the paper, they may have felt some slight embarrassment that there existed theorists in their field who still doubted the canonical understanding of gravitational radiation in general relativity.

6.4 Trading Zones and Pidgins

In his book *Image and Logic* (Galison 1997), another pioneer of the careful micro-study of physicists in action argues that different groups of scientists, in particular experimental and theoretical physicists, often speak different technical languages and encounter difficulty in communicating with each other. He argues that, in such situations, physicists find it useful to develop a pidgin, a term used to describe a secondary language, formed usually from a mishmash of other languages, used to facilitate trade between different peoples. Galison describes the conceptual space between different groups of physicists as a trading zone and discusses the use of pidgins, which in his usage may refer to particular mathematical constructs designed

³A reference to Kennefick (2007), illustrating one of the problems faced by an oral historian who wishes to write books and continue doing oral histories!

to permit experimenters and theoreticians (let's say) to discuss and compare the predictions of the latter with the results of the former.

The binary pulsar is an interesting case to observe the possible need for trading zones, since it was a discovery by radio astronomers who had, otherwise, relatively little contact with relativists interested in gravitational waves. At the same time, their field had arisen alongside the broader culture of relativistic astrophysics, which was formed by a first contact between radio astronomers and relativists after the discovery of quasars. To what extent do we observe the need for a trading zone between experimenters and theorists in our particular story? Certainly there seem to be areas of physics in which theorists and experimenters talk to each other regularly and apparently freely, and it is certainly also true that when physicists, even from very different subject areas, converse, they speak a recognizable technical language which seems to be quite unconscious of boundaries. Indeed, for the physicist, the international, intersubject quality of physics speech is one of the defining experiences of being a physicist (no doubt the same may be true for scholars in other disciplines). Nevertheless there is some evidence, in the case of the binary pulsar story, supporting the model put forward by Galison. One promising way to understand how scientists deal with trading zones, when and if they occur, is through the notion of *interactional expertise*, a concept which describes the ability of someone to talk intelligibly and usefully to an expert about their field, even if they are not (yet) capable of working in that field, which would be full expertise (Collins et al. 2007). It may be that, even where physicists lack direct expertise to work in a neighboring field, they at least possess interactional expertise to talk with their fellow physicists in that field.

Let us begin with the discovery of the binary pulsar in 1974. The two astronomers involved, Joseph Taylor and Russell Hulse, both received educations fairly typical of astronomers of their generation in that they were educated primarily in physics (in fact Hulse was still a graduate student when he discovered the binary pulsar). In this context, particularly as the two men were working in radio astronomy, astronomy is conceived of as being more or less a subdiscipline of physics, albeit an unusually ancient one which still maintained a certain level of institutional independence. As such they took courses in general relativity, a subject within physics which was typically considered an optional higher level course, but one which might be especially relevant to those planning to specialize in astronomy. As radio astronomers interested in pulsars, relativity theory was clearly relevant to an understanding of the source of the signals they planned to study, but not nearly as relevant and routine as the physics of the electromagnetically based detectors and instruments they operated.

Accordingly Joe Taylor describes one of his first actions on discovering that he had a binary pulsar with a uniquely close orbit involving unprecedentedly intense gravitational interaction between the two components (interview, 2nd May, 2008):

We'd both taken the obligatory, or almost obligatory, relativity course in University, as part of our physics training, but neither one of us was very deeply into relativity. My wife was much amused when one day, this was when I was at the University of Massachusetts, of course, I said I don't have to teach today, I'm going to drive into Boston and visit the

Tech Coop. And I spent the day in the MIT bookstore and came back with a pile of books, Weinberg, and Misner, Thorne and Wheeler and all the other ones that you would imagine. She was much amused that I spent the next few months deeply engrossed in these books.

So certainly the astronomers felt a need to get up to speed with the elements of relativistic orbital motion. To what extent was there a language gap between them and the practitioners of this discipline? Partly the gap was a social gap. Neither Taylor nor Hulse habituated among relativists and therefore did not partake in their discourse. So Taylor went unaware of the ongoing quadrupole formula controversy, throughout the time when, as we would be tempted to say today, he was determining the outcome of this controversy.

But leaving aside this question of discourse, when Taylor and his collaborators did speak to relativists, could they make themselves understood and be understood? Clearly they could, for the most part. But some obstacles were encountered. By the time Taylor and company were dealing with the orbital decay of the binary pulsar, Hulse had finished his doctorate and moved on. A collaborator with whom Taylor published many of the early papers announcing and discussing the orbital decay was Joel Weisberg. Weisberg does recall language difficulty playing some modest role in talking to theorists, before they found a long-term collaborator in a talented young French relativist, Thibault Damour (interview conducted by the author, by phone, on 24th February, 2000).

It's interesting, we had a failed attempt to work with one person. And I think the problem was he couldn't talk well enough to experimentalists. He couldn't give us results that were easily interpretable by us, whereas Thibault could. It was quite interesting.

Weisberg describes the kind of theorist that would be helpful in the process of theory testing using the binary pulsar data, saying "it had to be people who could talk a language I could understand." Regarding the one failed effort mentioned above, the problem had a very practical aspect, "he [the theorist] couldn't give us specific things to test." At the same time, he emphasizes that their eventual collaborator Damour was "brilliant" and "made fundamental progress," so "it wasn't just a language thing." He adds (in a private communication) that the "theorist 'speaking the right language' was not, by itself, enough for a successful collaboration."

Nevertheless, to examine the "language thing," I suspect it is fair to say that, in the absence of a relativity community, Taylor and Weisberg would have been capable of performing calculations to establish the predictions of certain theories. In fact, as we shall see, they did contribute original work on the theory side. The problem seems to me to be legitimately a question of language and society, in the sense that Taylor and Weisberg's problem was not primarily that they lacked the expertise to do the calculations. That much they could have acquired, and did acquire, with time and effort. What they lacked was fluency in the language spoken by theorists and social standing within the discourse of theory. The existence of theories to test is inextricably linked with the existence of theorists who developed them, who have a vested interest in the testing. Since the theorists are the experts, it is understandable that the astronomers, like Taylor and Weisberg, would feel distinctly hesitant about publicly putting forth calculations in an area that was not their own

realm of expertise. At the same time, it was important that the calculations which were done by theorists were not black boxes whose inner workings were totally opaque to the experimenters. It was important that the results of these calculations could be couched in a form which dealt with observables pertinent to the actual measurements being made.

The need for what Galison would describe as a pidgin seems to have produced the parameterized post-Newtonian (PPN) framework as a tool to mediate the theory-testing process. This process required an alliance of theorists and experimenters. Theorists made predictions based on their calculations. Experimenters made measurements which were then compared to the results of the calculations. This PPN framework had been widely used during solar system tests of general relativity, but was ill-adapted to the binary pulsar case because it presumed that the gravitational fields involved were very weak. Nevertheless a somewhat similar but much less general (focusing as it did upon the case of gravitational radiation emission) parameterization was established which facilitated the theory testing aspect of Weisberg and Taylor's 1981 paper. To quote from Clifford Will's paper on the subject (Will 1977):

Because of the complexity of many alternative theories of gravitation beyond the post-Newtonian approximation, we have not attempted to devise a general formulation analogous to the PPN framework beyond writing equation (2) with arbitrary parameters. However, we can provide a general description of the method used to arrive at equation (2), emphasizing those features that are common to the theories being studied.

So given the existence of a pidgin to create a trading zone between astronomers (and others) interested in doing theory testing and gravitational theorists, why did the astronomers shrink from commenting directly on the quadrupole formula itself? One obvious answer is that the pidgin was not designed to facilitate such a conversation. It permitted comparisons between calculations derived from different theories. It was not designed for the more complex and open-ended task of critiquing subtle details of such calculations. Another answer is that the barriers were as much social as linguistic (the two must obviously be linked). The astronomers felt they lacked the social standing to weigh in on a question which obviously fell within the purview of the theorists. Because the controversy over which calculation within a given theory was the correct one depended on subtle judgments, it naturally required the expertise of the practicing theorists. This is precisely the meaning of the theoreticians' regress that it depends on subtleties of expert judgment and not on some closed algorithmic model of how to perform a calculation.

6.5 Skeptics' Dilemma

I have argued that the closing of debate in the quadrupole formula controversy occurred at least partly because of the quickening effect caused by the binary pulsar increasing the importance of the controversy. At the same time, the lifetime of the

controversy, once the binary pulsar data became available, was greatly constrained by the existence of experimental data which bore directly on the topic at issue. For the theoretical controversy to continue indefinitely, there would have to have been a significant effort to contest either the experimental evidence or the interpretation of it. The fact that there was no such significant attack on the ruling interpretation of the binary pulsar data certainly limited the lifetime of the controversy, so it is interesting to look at the reaction of the skeptics to the work of Taylor and his collaborators.

In any problem of orbital mechanics, there are many mechanisms which might account for all or part of an observed change in orbital period. That even the most famous agreements between theory and observation can be challenged in this way is shown by the saga of Robert Dicke's efforts to measure the oblateness of the Sun (the degree to which its shape departs from a perfect sphere). Dicke had pointed out that if the solar oblateness turned out to be significantly different from zero, its gravitational influence on the orbit of Mercury would throw out the close agreement between the prediction of general relativity and the observed perihelion advance of the planet Mercury (Dicke and Goldenberg 1967). As with the case of the Mercury perihelion, the binary pulsar data seemed particularly impressive because it agreed with the prediction of the quadrupole formula with little or no need to take into account of other factors. The interpretation was that the system was very "clean." The corollary to this, naturally, is that any evidence that the system was not so clean would throw out the agreement. Given this opening to challenge the *interpretation* of the binary pulsar data, it is interesting that the gravitational wave skeptics were not involved in proposing alternative mechanisms.

Certainly there were those who considered it, among them Peter Havas and, very likely, his former student Arnold Rosenblum. They were to the fore in demanding that the observations not be accounted a successful test of general relativity given that (in their opinion) the quadrupole formula had not been shown to a valid prediction of that theory. Joe Taylor recalls that certain people were particular about this question of terminology (interview, 2nd May, 2008).

Well, let me think, the people who kept bugging me about it, so to speak, were Peter Havas, Fred Cooperstock and Arnold Rosenblum. Arnold bugged me about it a lot. Anyway, they just kept saying 'Look, even though you have an experimental number now, we're not even sure what the theoretical number is and you can't go around saying that you've confirmed something.' So I tried to remain outside of the argument, letting the theorists fight it out until they all . . . persuaded one another. So that seemed to be the best thing for me to do and we were simply concerned with getting an experimental result that we were happy with.

The alternative scenarios to the gravitational wave interpretation were actually put forward in print, but generally not by the skeptics. This may have been because the skeptics found themselves in a similar position to the experimenters. They had a vested interest in the debate, but lacked the special expertise which would have permitted them to comment. Likely dissipative mechanisms (or even non-dissipative ones) fell within the purview of astrophysics rather than relativity and were explored and commented upon by astrophysicists rather than relativists.

The most important issues which had to be dealt with in demonstrating that the observed decay agreed with the quadrupole formula prediction were the nature of

the unseen companion in the system and the relative acceleration of the binary pulsar to our solar system. If the unseen companion was a sufficiently compact object, like another neutron star (which is now firmly believed to be the case), then it would undergo little deformation as a result of the visible pulsar's tidal effect. But if it was a normal star, it would develop a marked oblateness which would in turn create a perturbation in the orbit of the pulsar (a tidal friction-like effect) which would be difficult, except over longer timescales, to distinguish from the orbital decay due to radiation damping. Effects of this type would, however, have affected other measurements made in the system, and with time the experimenters became convinced that the system was extraordinarily clean. As Taylor and McCulloch stated in their paper from the Texas Symposium (Taylor and McCulloch 1980):

If one were given the task of designing an ideal machine for testing gravitation theories, the result might be a system rather similar to PSR1913+16; an accurate clock of large mass and small size, moving at high speed in an eccentric orbit around a similar object located in otherwise empty space. To be sure, one would place the system somewhat closer to the Earth than ~ 5 kpc, or which arrange for a more powerful transmitter to convey the clock pulses to terrestrial telescopes; but we cannot expect Nature to be concerned with the inadequacies of our instrumentation!

This sense of wonder at the sheer serendipity of coming across such a system (many relativity theorists had sworn for decades that no system would ever be found in which gravitational wave effects would be measurable) was brought into focus for me after the more recent discovery of the "double pulsar," a system with an even closer orbit than the original binary pulsar, in which both pulsars are visible from Earth. I have heard this system referred to as "a relativistic astrophysicist's wet dream."

Taylor and McCulloch's comment illustrates the three main technical challenges in creating a match between theory and experiment for this system. First, the system must be in empty space. The presence of interstellar gas, for instance, would certainly alter the orbit of the system with time, as a result of dynamical friction. A related issue would be if the pulsars themselves were blowing off material at a significant rate, in which case the mass loss would affect the orbital motion. Secondly, as we have seen, both objects must be compact objects, such as neutron stars, so that perturbations due to the failure of the bodies to behave as point sources can be ignored. As a corollary to this, if the system contained a third massive object, this would obviously also affect the orbit of the two known components. Finally, the object should be close to us, not only for reasons of detection but because a more distant object is in a more different orbit around the center of the galaxy and would be accelerating more strongly with respect to us here on Earth (for a list of references and discussion of a number of these issues, see Damour and Taylor (1991).

It is a well-known result of special relativity that systems which are in inertial motion with respect to each other have clocks which run at different rates. If the systems are accelerating with respect to each other, then their respective clocks will alter, with time, in their relative rates of running. Since the solar system and the binary pulsar system are in different orbits around the galactic center, they are not in the same inertial frame with each other. Accordingly the sensitive timing

which is required to measure the orbital damping effect is also capable of measuring the relative accelerations of these two systems. In so far as doubt persisted about the validity of the quadrupole formula, this was a bad thing. Indeed, at one point during the 1980s, it did happen that the analysis of measurements of the binary pulsar did fall out of agreement with the quadrupole formula, by a much smaller amount than had been at issue in the earlier theoretical debate (in so far as that debate had ever been completely quantified). A close analysis of the relativistic theory of timing between the two systems, carried out by Taylor in collaboration with Thibault Damour, showed that the discrepancy could be explained on the basis of fully accounting for the timing issues (Damour and Taylor 1991).

Ultimately, as Taylor recalls, the situation reached the point where, if one *assumed* the validity of the quadrupole formula, one could make an accurate determination of the position of the binary pulsar in the galaxy, based on its relative acceleration. This measurement was more accurate than was possible by other methods at that time. This makes as good a moment as any to mark the end of the quadrupole formula controversy. When a prediction turns from a thing to be tested to a tool to be used, the debate is surely closed (and this, of course, goes some way to explain the impatience of non-skeptics to achieve that moment of closure). It is a mark of the importance of the controversy that the measurement of the distance to the galactic center which could have been provided by the binary pulsar data never became a canonical one, though it is in agreement with subsequent measurements using other techniques.

As Damour and Taylor put it in (Damour and Taylor 1991):

If we assume that the standard general relativistic framework . . . is valid we see that, in a few years, the measurement of \dot{P}_b^{obs} [the rate of decay of the binary pulsar's orbit] can be turned into a measurement of . . . the galactic constants R_o [the distance from the Solar System to the Galactic center] and v_o [the speed of galactic rotation at about the center at the position of the solar system] (especially v_o , which presently contributes the biggest uncertainty). Such a "pulsar timing" measurement of v_o would be free from many of the astrophysical uncertainties that have plagued other determinations.

Since the Taylor-Hulse discovery, subsequent binary pulsars have been found where the relative acceleration of the two systems does not permit a particularly accurate determination of the rate of orbital damping. Had the controversy persisted so far, this might have provided some opening for skeptics. However the discovery of the double pulsar in 2003, a system in which both pulsars are oriented so that both their radio beams are visible from the Earth, has provided a system with even stronger orbital damping than the original binary pulsars, whose results are in agreement with it.

How much interpretive flexibility was there for skeptics to continue the controversy? Did the skeptics largely abandon the fight because, as Franklin would have it, they were rational actors or, as Collins would have it, they had run out of sociological space in which to continue the argument? I suspect both considerations played a role. A rational actor will certainly take sociological considerations into account when determining whether to continue a debate. Most physicists do not

wish to face social ostracism, even in a cause they believe to be right. At the same time, any social constructivist will agree that the ruling out of certain arguments as work in the field progresses, the limitations placed on interpretative flexibility in the ebb and flow of debate, can tax the ingenuity of even the most stubborn skeptics to the point at which they give up the struggle. The social struggle can become unequal in a double sense, in that they are both outnumbered and outmaneuvered by their opponents. Whether the maneuvering was all in vain, given the inevitable verdict of nature is, of course, an interesting question, but not one that is trivial to answer by the historian's method.

That skeptics *considered* continuing the battle is clear enough. Although Fred Cooperstock did retire from the fray for a decade or so after the mid-1980s, he subsequently put forward a new argument that gravitational waves would not propagate energy through empty space (Cooperstock 1992). The fact that the first generation of gravitational wave detectors like LIGO failed to detect gravitational waves passing by the Earth provided a new opening for skeptics like Cooperstock. He and others now put forward arguments that the existing theory is correct for *sources* like the binary pulsar, but fails for *detectors* like LIGO, thus explaining why we see evidence for gravitational waves in these source systems, but cannot, as yet, detect them.⁴ The specifics of these new skeptical arguments vary widely. Some come from professional physicists like Cooperstock, others come from amateurs who focus on the sheer expense of the detectors which, they claim, can never succeed in detecting anything.⁵ In light of the new era of gravitational wave astronomy, at least some of these arguments now seem to be invalidated.

Peter Havas, when I interviewed him in 1995, certainly spoke of the openings he believed had existed, at least for a time, for an attack on the standard interpretation of the pulsar timing results. He still entertained significant doubts about the consensus which had emerged at that time. Joe Taylor reports that Havas, and his student Arnold Rosenblum, did ask to see some of the data and that he sent them a magnetic tape containing some (private communication). When he asked them a year later whether they had made progress, they indicated that they had been distracted by other problems. Nevertheless, a search for Arnold Rosenblum's papers on the SAO/NASA Astrophysics Data System server shows that, from the mid-1980s, after several years spent on his calculations of gravitational wave emission that did not agree with the quadrupole formula, he then devoted a number of papers to the problems of relativistic timing in orbital and binary systems. Although none of this

⁴We cannot hope, with current technology, to detect the gravitational waves emitted by the known binary pulsar systems. It is only when such systems reach their terminal point and spiral into each other and merge that Earth-based detectors can hope to observe them.

⁵A sample of modern gravitational wave skepticism is given by the following references: Cooperstock (1992), Bel (1996), Aldrovandi et al. (2008), and, for the non-professional viewpoint, see the webpage <http://www.god-does-not-play-dice.net/Szabados.html#SBG>, accessed on April 24th, 2009.

series of papers referred directly to the binary pulsar, they are strongly suggestive that he had spent a considerable amount of time thinking about this issue, leading him into that field.⁶

Therefore we can say that the skeptics considered a foray against the conventional interpretation of the binary pulsar data, but decided against it. One can say that the physics of the situation obliged them to react this way, in that they felt they could not overturn the hard empirical evidence provided by the binary pulsar data. But one can also say there were sociological reasons. They were not in a position to do their own experiment to challenge the data, because they lacked the standing in that field which would have permitted them to enter it with any hope of success. For starters they would never have been granted time on a radio telescope to do their own measurements of this system (one group of astronomers did do some independent timing measurements of the binary pulsar, guided by data supplied by Taylor, and concluded that Taylor and his collaborators were correct in their results on the orbital decay; see Boriakoff et al. (1982). Even worse, in so far as the interpretation of the data could be challenged by theorists, it was by astrophysicists with experience in the study of stellar binaries and pulsars, not by relativists experienced in gravitational waves. Thus, from a professional point of view, the skeptics were in a double bind which, combined with their increasing isolation within their own community, as the debate moved toward a final resolution, prevented any kind of continuation of the public debate. Whatever private doubts were held by a few theorists about the reliability of the existing calculations, the empirical result was regarded as beyond dispute. The final option open to the skeptics, arguing that Taylor had simply got it wrong, was undoubtedly not entertained because of the outstanding reputation which Taylor enjoyed within the astrophysics community for his careful and painstaking work.

This whole issue of replicating experimental work lies at the heart of the philosophical controversy between Collins and Franklin alluded to earlier. In his papers on the Weber controversy, Collins showed how problematic was the use of replication to close a debate. The difficulty lies in the fact that the details of experiments must necessarily differ, and these differences generally provide ammunition for one side or the other. In addition, what is sauce for the goose is sauce for the gander. If someone who replicates an experiment charges that the original experimenter got it wrong, the same charge can always be thrown back in their own teeth. Franklin responded that physicists could still rationally and (hopefully) dispassionately decide between these competing claims. From the sociologists' point of view, the issue casts interesting light on the nature of *expertise* and how it is recognized and evaluated by fellow experts. From the sociologists' standpoint, the physicist, as a rational actor, must make a series of social judgments over the course of a controversy in evaluating his colleagues' expertise and the consequent reliability of their work.

⁶Arnold Rosenblum died tragically at a young age in 1991 (Cohen et al. 1991).

In the case of the binary pulsar, replication demanded access to radio telescope time to look at the same system or, better, the discovery of an independent system. But, as we have seen, subsequent systems were often not as ideal for this experiment as the original. Not until the discovery of the double pulsar can we be said to have a fully comparable replication of the original, so one can certainly speculate that there may have been some scope for further controversy in the decades between 1980 and the early years of the twentieth century, had there been sufficient sociological space to support such a debate. But while logical space for disputation may have remained, the skeptics had run out of sociological space. Indeed, there is every reason to believe that the field of gravitational wave physics could ill afford to permit such a controversy to linger for that amount of time, lest it put its own disciplinary space at risk.

6.6 Theory Testing

An interesting feature of the early papers on the orbital decay measurements of the binary pulsar is the focus on theory testing. In the 1981 paper by Weisberg and Taylor, much of the paper is devoted to discussion of the predictions of a variety of alternative theories of gravity which were falsified by the measurements. The best known of these theories was the bi-metric theory of Nathan Rosen, a longtime skeptic of gravitational waves (Rosen 1940). Rosen's theory had been shown to make a prediction of anti-damping for binaries emitting gravitational waves (Will and Eardley 1977). The waves would carry away negative energy from the system, leaving it more energetic than before and thus permitting the orbiting bodies to spiral away from each other. As Will and Eardley acknowledged in Will and Eardley (1977, p. 92), "some might thereby argue that the theory should be ruled out on theoretical grounds alone." But the theory was of particular interest to theory testers because it agreed with the predictions of general relativity in the post-Newtonian limit. Accordingly it was one of a handful of theories which had survived all early solar system tests of gravitation theories. This placed it in a special category of theories which could play a useful role as a foil for theory testing with the binary pulsar, even if calculations such as these were beginnings to show that the theory had troubling pathologies.

The emphasis on theory testing in early papers on the orbital decay seems odd when these papers are read today. This work is famous, but certainly not famous for invalidating the bi-metric theory of Nathan Rosen. What seems particularly odd is that the prediction of Rosen's theory (and other theories) which it invalidated appeared paradoxical. As was openly acknowledged, no one would give any credence to these predictions even without an experiment to falsify them. The theory had few, if any, proponents by this stage. It's certainly true that we cannot say here that we have a direct confrontation between theories, in any symmetric sense. While Rosen's theory could have been falsified by the results, it could not have been confirmed. Had Taylor and colleagues encountered a result in agreement

with Rosen's theory's prediction, all sorts of other mechanisms would have been proposed to explain it before Rosen's. Perhaps, given enough supporting evidence (several other systems behaving the same way), Rosen's theory would have been accepted, but it would have been a long hard road.

Nevertheless, no matter how little credibility a theory has, experimenters still find it satisfying to have a definite prediction they can test. One must not be too inclined to overlook the obvious motivation. Indeed, a crazy prediction of a reasonable theory, as long as it is a definite prediction, may be a godsend to an experimentalist. After all, falsifying such a prediction is likely to be seen as good, worthwhile work by colleagues, and yet it will also be uncontroversial and easily accepted by the community. Furthermore a theory like Rosen's, with its odd prediction, plays a useful role in the framing of experimental results as theory testing. It is a straw house theory, in the sense that it is rather like the pig who built his house out of straw. The main purpose of the research is to show that general relativity has been validated. Therefore general relativity is like the house of bricks which does not fall down to the huffing and puffing of the big bad wolf. But the story of the one little pig is rarely satisfying to an audience. In order to appreciate the part about the pig who survived, we must first learn about his brothers who were not so lucky. The foolish pigs who built their houses of straw and sticks are perhaps all the more welcome, from a narrative standpoint, if the brick house is the subject of controversy. Doubts have been voiced as to whether the brick house really was built by the third pig. Perhaps, say his detractors, said pig has been given too much credit. How much easier it is to talk about the first two pigs. At least no one is trying to claim credit for their edifices!⁷

Lest anyone think that it is normal to find theories with strange predictions waiting to be falsified, one must give Rosen some due credit here for not contesting the calculation which set up this scenario. Clifford Will, a leading figure among theorists interested in theory-testing experiments and the chief architect of the parameterized post-Newtonian scheme mentioned previously, was very active in producing the calculations which provided predictions from alternative theories. Note the very fact that the authors of the theories were not doing these calculations themselves suggests that we are not dealing with theories which have proper communities of advocates behind them. Will notes that it is relatively unusual to find that the author of a new theory will agree with a calculation which shows that the theory makes a prediction that is highly likely to be falsified by experiment. Generally the process, typically of many scientific controversies, can be almost open-ended (Interview conducted by the author at Washington University, St. Louis, 2nd March, 1999):

It can be, and it rarely reaches a conclusion. The only time I know . . . and I don't get involved in this all that much. I mean, I don't grab theories out of the literature and analyze them. It's kind of a hopeless and not a terribly rewarding task. But my experience has been that it takes a long time because the people who propose it always try to wriggle out of it. But there are only two cases that I know of where it has actually come to a conclusion whereby the

⁷Parts of this section are based on an unpublished paper by the author and Harry Collins.

person said, 'yes I agree, this theory is wrong' and one was Rosen himself. Because when we did this work on the binary pulsar and showed that Rosen's theory disagreed with the observations, in fact I was giving a talk in Haifa shortly after that and gave this lecture and said that Rosen's theory is wrong and at the end of the lecture Rosen stood up and said 'yep, I agree with you, it's wrong, . . . but I have a new theory' a rather different theory which he then went on to argue had nice properties and agreed with all the experiments.

It is worth noting here that Will's work showed that Rosen's theory predicted negative energy wave emission only in the case of dipole gravitational waves. This in itself was a departure from standard general relativity theory, since dipole gravitational waves do not exist in this theory. Even in Rosen's theory, dipole radiation would not be emitted for a binary system consisting of two identical pulsars. Since it was gradually shown that the binary pulsar and its companion are fairly similar, Weisberg and Taylor (1981) found it necessary to calculate the quadrupole prediction of Rosen's theory (and certain other theories which were also falsified by their work) themselves, drawing upon Will's framework. The calculation showed that Rosen's theory predicted negative energy waves even in the quadrupole case.

But if Rosen's theory predicted an unphysical result, wouldn't it have been discarded even if the binary pulsar hadn't been found to falsify it? To quote Cliff Will again (interview, 2nd March, 1999):

In a case like that it really depends on your point of view. Some people would have argued that just having anti-damping, negative energy flux would make that a bad theory right off the bat and you would throw it away without further ado. So my attitude is slightly more phenomenological than that. I'm willing to say that it looks strange to me, but let's compare it with observations and, of course, there the comparison is easy because we see damping and not anti-damping and so it really is wiped out. But some people would just say on theoretical grounds, 'that theory's dead.'

The context here is particularly important. Physicists interested in gravitational waves were used to having no experiments at all. Once an experiment had, at last, become available, they wanted to put it to every kind of use they could, and theory testing was the most established role for experiment in the general relativity community. Most of the work in this field which had some prestige in the wider physics community was of the theory-testing variety, such as the British 1919 eclipse expedition, the Pound-Rebka experiment, the perihelion advance of Mercury, and the Shapiro time-delay measurement. The limitation of all of these experiments, as far as theory-testing goes, was that they were all "solar system" tests limited to weak gravitational fields. As such, some theories, and Rosen's was a leading example, could not be distinguished from general relativity by these tests. It was certainly natural for those involved with the binary pulsar to anticipate that its significance would lie largely in the fact that it was the first strong-field test of general relativity and its rivals. In fact, such was the significance of the discovery of evidence of the existence of gravitational waves that this quickly came to dominate everything else. As such, the falsification of Rosen's theory seems almost quaint today, compared to the importance of the verification of the general relativistic quadrupole formula.

6.7 Conclusions

It is now time for me to examine my own place within a controversial field, in analogy with my study of the astrophysicists struggling to interpret the binary pulsar data. I find myself trying to interpret their struggles in the context of the competing theories of social constructivism and rival philosophies which insist on the normative standing of experiment, permitting it a special status in deciding scientific controversies. It is in this sense that the binary pulsar story may be said to be a potentially controversial case study from the science studies standpoint.

Is there a sense in which my work can decide between these competing theories? Unfortunately the answer appears to be no. Perhaps this is fortunate, since I mentioned at the outside that I am not sure I want to place myself squarely in the cross hairs of this particular controversy. The problem is that the predictions of the two theories do not significantly differ from each other in this case. At the resolution provided by my study, there does not seem to exist a possibility of deciding between them. Collins would say that Kennefick has extended his notion of the experimenters' regress onto the theoretical side in a way that seems natural and useful. He would say it is not at all surprising that the theorists' seeking a way out of the regress should turn to an outside expertise, in the form of experimenters, to find a resolution. It is just the inverse of the way in which experimenters, seeking a way out of their regress, might appeal to theory in order to decide between competing experimental results. But of course Franklin is perfectly happy for theorists to let their debates be decided by experiment. It fits in completely with his normative picture of experiment as the decisive factor in such disputes. Thus each side is likely to be happy with the basic story I've outlined. Even more problematic is the way in which the philosophical debate does not necessarily permit a clear distinction to be drawn in the behavior of the scientists involved. Even if we concede that the protagonists were more willing to settle the issue based on what the experimenters said and were relatively unwilling to challenge what the experimenters said, this could be explained by the sociologists as simply a feature of the society under examination. Theorists, by the rules of the game, have a relatively limited (but definitely non-zero) liberty to challenge the expertise of experimenters. When they do so, they must do so from a position of strength, and the entire history of the controversy shows that the skeptics were already in a position of weakness by the time the binary pulsar data came along. Indeed we have noted that a paradoxical effect of the experimenters' arrival on the scene was to breathe new life into the controversy, effectively giving new oxygen to the skeptics, even as it forced them to consume more oxygen in the exertion of defending their position.

Recall that the substance of the debate between the physicist/philosopher Franklin and the sociologist Collins was the problem of replication and how one can tell whether an experiment has been performed correctly. To some extent it boils down to the question of how scientists deal with the possibility that Joe Taylor and colleagues might simply have gotten it wrong. This is especially noteworthy in this case, because for a considerable time there was no confirming experiment.

The answer is that most people were impressed with Taylor and felt they had every reason to trust his work. This is a profoundly sociological issue obviously. Even if one believes that Taylor is correct, one has not actually done all the work he has done to convince himself. One assumes that he has done a proficient job, especially if one has had reason to believe that other work he has done has been very reliable. In short, at least one important aspect of judging the work of fellow physicists derives from our ability to judge their standing in the community and to assess their expertise from social encounters. Note that only if we are physicists ourselves are we likely to have much success in making this kind of judgment. Collins' and Franklin's debate concerns (in part) the question of how much importance one should place on this aspect of the reception of scientific work.

For what it is worth, my own view is that at bottom physicists are simply doing what humans usually do and applying a basic version of the principle of induction. If an experiment is replicated n times and always produces the same result, then the $n + 1$ th replication will produce the same result. It makes sense for physicists to infer that this is because reality is determining the outcome of the experiment. For a sociologist, it makes more sense to assume that if n experimenters sharing a similar expertise perform an experiment the same way, then the $n + 1$ th expert will perform it the same way and also produce the same result. The inference is different, depending on the academic interests of the scholars involved, but the basic principle is the same. It seems to me that the argument between some sociologists and some philosophers on this topic is similar to the old dispute between realists and empiricists. Philosophers are saying that science is only possible because scientists are engaged in studying a real entity, the laws of nature. In this case, the sociologists are in the role of the empiricists, insisting that sense impressions are the only reality and observing that much of what passes for sense perception in modern science is what scientists hear from other scientists as scientific knowledge passes through a series of social networks. I doubt that I can decide a debate between realism and empiricism.⁸

The moral of this story, it now seems to me, is that science works. Does this mean that the story I am telling is an argument in favor of realism? It's certainly not an argument against realism, and it is true that the strong realist would say that

⁸Another question I cannot answer is if n historians study the same historical episode, can we rely on the $n + 1$ th historian reaching the same conclusions? Can I do so if $n = 1$? Is there any sense in which historical micro-studies of this kind can be compared to real science? Is there a historians' regress related to the problem of when history ends, just as the experimenters' regress relates to the problem of when experiments end? The phrase "when history ends" may seem millennial in tone, but note the aptness of the word apocalypse which means "the lifting of the veil," which is certainly what the historian is trying to do. Just as the radio astronomer does in continuing his timing measurements over longer periods to greater degrees of precision, or as the theorist does in delving to higher orders in an approximation scheme, so the historian burrows down more deeply in a micro-study. But in historical analysis we should be careful to practice Interpretational Frugality, a sort of inverse form of Occam's Razor. We may multiply entities if it is in the service of keeping our feet grounded in the local. Not all morals are generally applicable. In the words of Bart Simpson, sometimes there is no moral, "just a bunch of stuff that happened."

science has to work because the objective nature of reality constantly obtrudes on experimental work of all kinds. But if we adopt the position of the strong program of the sociologists, we must work rather harder to explain how it is that scientists “manufacture consent.” This question is of interest even to the realist, since history certainly tells us that people are sometimes wrong about the laws of nature. In this imperfect world, if scientists are able to reach agreement among themselves, we can announce that science works. This sounds like a very global moral, but its true significance is local. The relativity community, I believe, had quite a lot at stake in this debate. They had to show that *their* science worked and that they as a community could do science which worked. The binary pulsar therefore played a key role in showing that relativity as a field, and relativists as a community, could work as a functioning branch of science and that relativists were competent and not dopes. From the point of view of the relativists, it would have mattered little whether their behavior was viewed as that of incompetent, irrational physicists who refused to accept the obvious fact that gravitational waves existed or as that of needlessly fractious and insufficiently socialized actors unable to crystallize a core group among themselves in order to facilitate normal scientific behavior. What they had to do was demonstrate that they were a healthy branch of physics, a postulate which Feynman, sitting in the Grand Hotel, Warsaw, in 1962, would have doubted. The imagery confronting Feynman as he set in a hotel restaurant and contemplated this dysfunctional field, writing the script for a possible Fellini film, “126 Dopes,” has been replaced by the gleaming, high-tech, ultra-precise big science of LIGO and by the confidence of funders in pouring money into the construction of large gravitational wave detectors. For a decade no one knew for certain whether gravitational waves would be detected by these devices, so the process by which society decided to begin ignoring the anxieties of the skeptics is an interesting one regardless of whether we believe those cautious skeptics were irrational dopes or sensible social actors.

Now a second potentially Nobel Prize winning experimental result seems to have eliminated all doubts of the existence of gravitational waves. Here is where Collins and Franklin both become excited. Franklin says that the view that gravitational waves are real has been completely vindicated and the era of gravitational wave astronomy can begin. Mission accomplished for one task facing physicists. Now they can move onto the next technical challenge. Collins says a new form of reality has entered the social world. Where once we believed merely in artifacts in the detector’s output, now we believe that enormous events across the Universe are part of our social reality here on Earth. Now we move onto the next sociological challenge facing gravitational wave physicists. How to adapt to the sociological mores of the field of astronomy, from those found in the field of high-precision physics experiments? As one example of this, up to now they have refused to announce detections that fall below a certain threshold in terms of the signal-to-noise ratio in the instruments. The reason is because they have been worried that they could be wrong about a less statistically certain signal and people might refuse to believe in it. Now that people are satisfied about the reality of gravitational waves being emitted by binary black holes, astronomers are interested in the vital question

of how many black holes are out there and what their masses are. To refuse to accept signals falling below a high threshold is to willfully ignore quite likely detections and therefore to bias the data toward larger and closer black hole binaries. Thus clinging to the standards of high-precision physics could compromise the science of astrophysics. From a sociologist's viewpoint, it will be fascinating to see how the physicists grapple with this change of culture. From the physicist's viewpoint, it will be fascinating to see what they discover.

Acknowledgements I would like to thank Joseph Taylor, Clifford Will, Thibault Damour, Joel Weisberg, and the late Peter Havas, all of whom permitted me to interview them for the research which gave rise to this paper. All of the interviews, except the one with Peter Havas, were recorded. Both Harry Collins and Allan Franklin discussed some of the issues bearing on this paper with me many times, and aspects of it are based on an unpublished draft of a paper written by Collins and me. I would like to thank both of them for their help and inspiration on this work. Diana Buchwald and Kip Thorne both helped me far more than I can recall in the early stages of this work, and I would also like to thank David Rowe for giving me the chance to finally turn it into a paper and for his patience waiting for it to be finished.

References

- Aldrovandi, R., Pereira, J. G., da Rocha, R., & Vu, H. K. (2008). Nonlinear gravitational waves: Their form and effects. arXiv:0809.2911v1.
- Baade, W., & Zwicky, F. (1933). Remarks on super-novae and cosmic rays. *Physical Review*, *46*, 76–77.
- Bel, L. (1996). Static elastic deformations in general relativity. electronic preprint gr-qc/9609045 from the archive <http://xxx.lanl.gov>.
- Boriakoff, V., Ferguson, D. C., Haugan, M. P., Terzian, Y., & Teukolsky, S. A. (1982). Timing observations of the binary pulsar PSR 1913+16. *The Astrophysical Journal*, *261*, L97-L101.
- Cohen, J. M., Havas, P., & Lind, V. G. (1991). Arnold Rosenblum. *Physics Today*, *45*, 81. Another obituary of Rosenblum appeared in the *New York Times* of January 7, 1991.
- Collins, H. M. (1994). A strong confirmation of the experimenters' regress. *Studies in History and Philosophy of Science Part A*, *25*, 493–503.
- Collins, H. M. (2004). *Gravity's shadow*. Chicago: University of Chicago Press.
- Collins, H. M. (2009). We cannot live by scepticism alone. *Nature*, *458*, 30–31.
- Collins, H. M., Evans, R., & Gorman, M. (2007). Trading zones and interactional expertise. *Studies in the History and Philosophy of Science A*, *38*, 657–666.
- Cooperstock, F. I. (1992). Energy localization in general relativity: A new hypothesis. *Foundations of Physics*, *22*, 1011–1024.
- Damour, T., & Ruffini, R. (1974). Sur certaines vérifications nouvelles de la Relativité générale rendues possibles par la découverte d'un pulsar membre d'un système binaire. *Comptes Rendu de l'Académie des Sciences de Paris, series A*, *279*, 971–973.
- Damour, T., & Taylor, J. H. (1991). On the orbital period change of the binary pulsar PSR 1913+16. *The Astrophysical Journal*, *366*, 501–511.
- De Witt, C. M. (1957). *Conference on the role of gravitation in physics*, proceedings of conference at Chapel Hill, North Carolina, January 18–23, 1957 (Wright Air Development Center (WADC) technical report 57–216, United States Air Force, Wright-Patterson Air Force Base, Ohio). A supplement with an expanded synopsis of Feynman's remarks was also distributed to participants (a copy can be found, for example, in the Feynman papers at Caltech). This report has recently been published online by Edition Open Sources and can be accessed at <http://www.edition-open-sources.org/sources/5/index.html>.

- Dicke, R. H., & Goldenberg, H. M. (1967). Solar oblateness and general relativity. *Physical Review Letters*, 18, 313–316.
- Dyson, F. (1963). Gravitational machines. In A. G. W. Cameron (Ed.), *Interstellar communications* (pp. 115–120). New York: W.A. Benjamin Inc.
- Einstein, A. (1916). Näherungsweise Integration der Feldgleichungen der Gravitation. *Königlich Preussische Akademie der Wissenschaften Berlin, Sitzungsberichte* (pp. 688–696)
- Einstein, A. (1918). Über Gravitationswellen. *Königlich Preussische Akademie der Wissenschaften Berlin, Sitzungsberichte* (pp. 154–167)
- Einstein, A., & Rosen, N. (1937). On gravitational waves. *Journal of the Franklin Institute*, 223, 43–54.
- Feynman, R. P., & Leighton, R. (1988). *What do you care what other people think? Further adventures of a curious character*. New York: Norton. Remark quoted appears on pg. 91 of the Bantam paperback edition (New York, 1989).
- Franklin, A. (1994). How to avoid the experimenters' regress. *Studies in History and Philosophy of Science Part A*, 25, 463–491.
- Galison, P. (1997). *Image and logic: A material culture of microphysics*. Chicago: University of Chicago Press.
- Haensel, P., Potekhin, A. Y., & Yakovlev, D. G. (2007). *Neutron stars 1: Equation of state and structure*. New York: Springer.
- Havas, P. (1973). Equations of motion, radiation reaction, and gravitational radiation. In *Ondes et Radiation Gravitationnelles* proceedings of meeting, Paris, June, 1973 Paris: Editions du Centre National de la recherche scientifique (pp. 383–392).
- Hulse, R. (1997). The discovery of the binary pulsar. In G. Ekspong (Ed.), *Nobel Lectures in Physics 1991–1995*. Singapore: World Scientific.
- Hulse, R. A., & Taylor, J. H. (1975). Discovery of a pulsar in a binary system. *Astrophysical Journal*, 195, L51–L53.
- Kennefick, D. (2007). *Traveling at the Speed of Thought: Einstein and the Quest for Gravitational Waves*. Princeton, NJ: Princeton University Press.
- Rosen, N. (1940). General relativity and flat space I. *Physical Review*, 57, 147–150.
- Royal Swedish Academy of Sciences. (1993). Press Release announcing the Nobel prize winners in Physics for 1993, issued 13 October, 1993 and retrieved on the web at http://nobelprize.org/nobel_prizes/physics/laureates/1993/press.html on Apr 21, 1993.
- Taylor, J. H., & McCulloch, P. M. (1980). Evidence for the existence of gravitational radiation from measurements of the binary pulsar 1913+16. In J. Ehlers, J. Perry, & M. Walker (Eds.), *Proceedings of the Ninth Texas Symposium on Relativistic Astrophysics* (pp. 442–446). New York: New York Academy of Sciences.
- Wagoner, R. V. (1975). Test for the existence of gravitational radiation. *Astrophysical Journal*, 196, L63–L65.
- Weisberg, J. M., & Taylor, J. H. (1981). Gravitational radiation from an orbiting pulsar. *General Relativity and Gravitation*, 13, 1–6.
- Will, C. M. (1977). Gravitational radiation from binary systems in alternative metric theories of gravity: Dipole radiation and the binary pulsar. *The Astrophysics Journal*, 214, 826–839.
- Will, C. M., & Eardley, D. M. (1977). Dipole gravitational radiation in Rosen's theory of gravity – Observable effects in the binary system PSR 1913+16. *The Astrophysical Journal*, 212, L91–L94.

Chapter 7

The Rise and Fall of the Fifth Force



Allan Franklin

On January 8, 1986, a headline in the *New York Times* announced, “Hints of Fifth Force¹ in Nature Challenge Galileo’s Findings.”² Four years later at the January 1990 Moriond Workshop,³ Orrin Fackler, one of the experimenters working on the Fifth Force, stated, “The Fifth Force is dead.” The workshop was attended by representatives of virtually every group then working on the Fifth Force. No one disagreed.

In this essay I will outline the short, happy life of the Fifth Force, a proposed modification of Newton’s law of universal gravitation, involving both the composition dependence and the distance dependence of the force, from its origins to its demise.⁴ The story begins with two seemingly independent strands: 1) K-meson decay and CP violation and 2) modifications of Newtonian gravity. When these two strands came together, the Fifth Force was born.

¹Physicists, at the time, spoke of four forces: 1) the strong or nuclear force, which holds the atomic nucleus together; 2) the electromagnetic force, which holds the atom together; 3) the weak force responsible for radioactive decay; and 4) gravity. Although the Fifth Force was a proposed modification of gravity, it involved the exchange of a different particle, a massive scalar particle, and so was considered as another force.

²This was a reference to the fact that the proposed Fifth Force, unlike gravity, was composition dependent. The Fifth Force between two lead masses would be different than the Fifth Force between a lead mass and a copper mass. The Fifth Force, as discussed below, also differed from the force of gravity in its dependence on the distance between the masses.

³The Moriond Workshops, devoted to “new and exotic phenomena,” were very important in the history of the Fifth Force. Not only were new results presented, but there was rigorous criticism of the new work, both formal and informal.

⁴For a more complete and detailed history, see Franklin (1993).

A. Franklin (✉)

Department of Physics, University of Colorado, Boulder, CO, USA

e-mail: allan.franklin@colorado.edu

7.1 The Rise ...

7.1.1 *K-Meson Decay and CP Violation*

The history of the Fifth Force begins with a seeming digression because it involved not a modification of gravitational theory but rather an experimental test of and confirmation of that theory. In 1975 Colella et al. (1975) measured the quantum mechanical phase difference between two neutron beams caused by a gravitational field. Although this experiment showed the effects of gravity at the quantum level, it did not distinguish between general relativity and its competitors. This was because the experiment was performed at low speeds, where, as Ephraim Fischbach pointed out, all existing gravitational theories predicted the same results (Fischbach and Freeman 1979; Fischbach 1980). In this work Fischbach also considered whether gravitational effects might explain the previously observed violation of CP symmetry (combined particle-antiparticle and parity or space-reflection symmetry) in K_L^0 decays.⁵ Fischbach pointed out that there were both experimental and theoretical arguments against gravity as the source of CP violation but wondered whether they were relevant to his work.

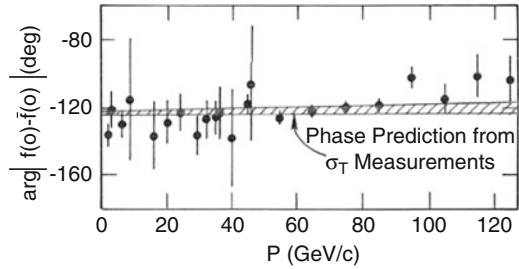
Theorists had already noted that for a long-range field that coupled differently to the K^0 and anti- K^0 mesons, a hyperphoton, and CP-violating effects would be proportional to the square of the K_L^0 energy (Bell and Perring 1964; Bernstein et al. 1964).⁶ Weinberg (1964) had also shown that because neither strangeness nor isotopic spin, the supposed sources of the field, was conserved, the K^0 mesons, as well as all strange particles, would be totally unstable if the range of the force was the size of our galaxy.⁷ (The ratio ($K_S^0 \rightarrow 2\pi + \text{hyperphoton}/K_S^0 \rightarrow 2\pi$) would be approximately 10^{19}). These issues became moot when experiments showed that CP violation was constant as a function of energy (Galbraith et al. 1965; DeBouard et al. 1965).

⁵CP symmetry allows the K_S^0 meson, the short-lived neutral K meson, but not the K_L^0 meson, its long-lived counterpart, to decay into two pions. In 1964 Fitch and Cronin and their collaborators (Christenson et al. 1964) found evidence for the two-pion decay for the K_L^0 meson and thus for CP violation.

⁶The K mesons, along with the Λ hyperon, had rather peculiar properties. They were copiously produced in strong interactions but decayed rather slowly by means of the weak interaction. No other particles, at the time, behaved in this manner. This led Gell-Mann and Nishijima to suggest that the K mesons possessed a property called strangeness, which was conserved in the strong, but not in the weak, interactions. This would explain the odd properties of the K mesons. The K^0 and its antiparticle the anti- K^0 had strangeness 1 and -1 , respectively. At the time of the Fifth Force, the conservation of strangeness was an established conservation law. When physicists spoke of the strong interactions, they spoke of the K^0 the anti- K^0 mesons. In discussing the weak interaction, they spoke of the K_S^0 and K_L^0 mesons, which were different linear combinations of the K^0 the anti- K^0 mesons.

⁷The K^0 mesons would be stable if the range of the force was of the order of the radius of the Earth, something Weinberg regarded as unlikely.

Fig. 7.1 The phase of the regeneration amplitude as a function of momentum. From Bock et al. (1979)



Fischbach was also encouraged by what he regarded as a “remarkable numerical relation.” Using his calculated energy scale for the gravitational effect, gh/c , Δm , the known $K_L^0 - K_S^0$ mass difference, and an enhancement factor of $m_K/\Delta m$, for which no justification was given, he found that the gravitational effect in CP violation was 0.844×10^{-3} , whereas the CP-violating parameter $1/2\text{Re}(\epsilon)$ was approximately equal to 0.82×10^{-3} . This seems indeed to be a remarkable coincidence because there is no known connection between gravity and CP violation. It is made even more remarkable when one realizes that the enhancement factor $m_K/\Delta m = 1.4 \times 10^{14}$.

A relativistic version of the experiment of Colella and colleagues did not seem feasible, so Fischbach began, in collaboration with Sam Aronson, an experimenter with considerable experience on K -meson experiments, to investigate whether such an experiment would be possible with K mesons. At this time, in the early 1980s, Aronson and his collaborators had been investigating the regeneration of K_S^0 mesons and found what seemed to be an energy dependence of the phase of the regeneration amplitude.⁸ Although the results were consistent with a constant phase, the low-energy points have a larger phase than the high-energy points (Figure 7.1). Further investigation by Aronson, Fischbach, and their collaborators (Aronson et al. 1983a,b) revealed several suggestive energy dependences in the CP-violating parameters. Figure 7.2 shows the most significant effect. They concluded that, “The experimental results quoted in this paper are of limited statistical significance. *The evidence of a positive effect in the energy dependencies of (the parameters) is extremely tantalizing, but not conclusive*” (Aronson et al. 1983a, p. 488).⁹ The experimenters concluded, “It is clear, however, that if the data... are correct, then the source of these effects will represent a new and hitherto unexplored realm of physics” (Aronson et al. 1983b, p. 516). An unkind referee remarked, “This latter statement also applies to spoon bending.”¹⁰ The paper was, however, published.

⁸The phenomenon of regeneration was one of the very unusual properties of the K^0 mesons. An accelerator-produced beam of K^0 mesons contains 50% K_S^0 mesons and 50% K_L^0 mesons. If one waited until all of the K_S^0 mesons decayed and then allowed the remaining K_L^0 mesons to interact with matter, one found that the beam once again contained K_S^0 mesons. They had been regenerated.

⁹These energy dependences later disappeared, but at the time, they were “tantalizing” effects.

¹⁰Ephraim Fischbach gave me a copy of the referee’s report.

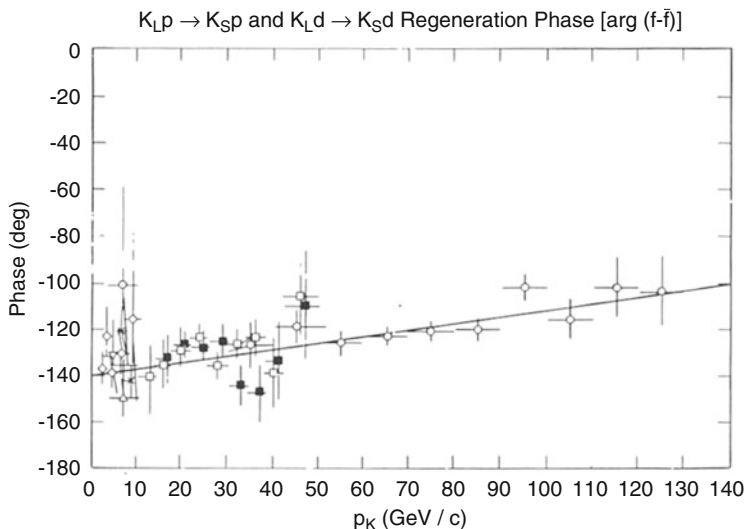


Fig. 7.2 The phase of the CP-violating amplitude as a function of momentum. From Aronson et al. (1983a).

7.1.2 Modifications of Newtonian Gravity

The second strand of our story involved proposed modifications of Newtonian gravity. Newtonian gravity and its successor, Einstein's general theory of relativity, although strongly supported by existing experimental evidence,¹¹ have not been without competitors. Thus, Brans and Dicke (1961) had offered a scalar-tensor alternative to general relativity. The theory contained a parameter ω , which for large values made the theory indistinguishable from general relativity. At this time ω had been found to be greater than 500, making the two theories indistinguishable.

In the early 1970s, Fujii (1971, 1972, 1974) suggested a modification of the Brans-Dicke theory that required a new, and hitherto unobserved, massive, scalar, exchange particle, in addition to the massless scalar and tensor particles of the Brans-Dicke theory. He found that including such a particle gave rise to an additional short-range force, of the order of 10 m–30 km, depending on details of the model. In Fujii's theory the gravitational potential took the form $V = -GmM/r[1 + \alpha e^{-r/\lambda}]$, where α was the strength of the new interaction and λ was its range. The first term was the ordinary gravitational potential. The second term was Fujii's modification. Fujii's model also predicted a gravitational constant

¹¹For an excellent and accessible discussion of this, see Will (1984). For more technical details, see Will (1981).

that varied with distance¹² and that the gravitational constant at large distances, G_∞ , would be equal to $3/4G_{\text{LAB}}$, the value at short distances.

Fujii also searched for possible experimental tests of his theory. Most interestingly for our story, he discussed the famous experimental test of Einstein's equivalence principle that had been performed by Roland von Eötvös and his collaborators in the early twentieth century and published in 1922 (Eötvös et al. 1922; this experiment, which is crucial to our history, is discussed below). Fujii noted that his new force predicted an effect that was smaller than the upper limit of five parts in 10^9 set by Eötvös, whose experiment was sensitive to such a short-range force. Fujii suggested redoing the Eötvös experiment and also other suggested possible geophysics experiments. He remarked that, although his calculated effect was, in fact, smaller than the limit Eötvös had set, local mass inhomogeneities would pose difficulties. As we shall see, this was a prescient comment.

Long (1974) investigated whether Newtonian gravity was valid at laboratory distances and found a small effect.¹³ Long's work led Mikkelsen and Newman (1977) to examine the status of G , the gravitational constant. They concluded, "Constraints on G in the intermediate distance range from $10\text{ m} < r < 1\text{ km}$ are so poor that one cannot rule out the possibility that $G_c[G_\infty]$ differs greatly from $G_0[G_{\text{LAB}}]$ " (Mikkelsen and Newman 1977, p. 919). They pointed out that their analysis "does not even rule out Fujii's suggested value $G_c/G_0 = 0.75$ " (Mikkelsen and Newman 1977, p. 924).

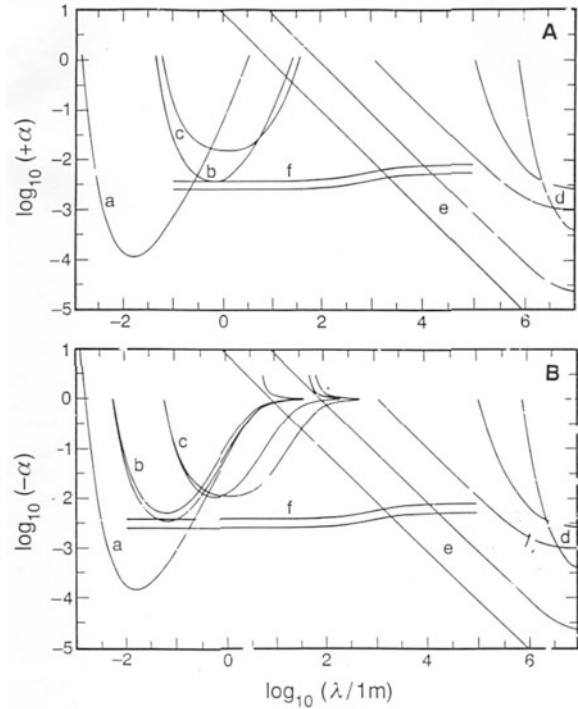
The most important summary of work on G , from the point of view of the subsequent history of the Fifth Force, was that given by Gibbons and Whiting (1981). Their survey included measurements of gravity in mineshafts and in submarines. The results for G from those measurements were slightly higher than those obtained in the laboratory, but because of experimental uncertainties, no firm conclusion could be drawn. Gibbons and Whiting summarized the situation as follows. "It has been argued that our experimental knowledge of gravitational forces between 1 m and 10 km is so poor that it allows a considerable difference between the laboratory measured gravitational constant and its value on astronomical scales, an effect predicted in theories of the type alluded to above [these included Fujii's theory]" (Gibbons and Whiting 1981, p. 636). Although experiment allowed for such a difference between the laboratory and astronomical values of G , there were reasonably stringent limits on any proposed modification in the distance range 1–10 km. There was, however, a small window of opportunity for a force with a strength approximately one percent that of gravity and with a range between 1 meter and 1 kilometer (Figure 7.3).

At this time there were also hints that the value of G measured in the laboratory differed from that found in geophysics experiments, although experimental uncertainties precluded a definite conclusion (Stacey and Tuck 1981; Stacey et al. 1981).

¹²Although a varying constant seems like an oxymoron, it is useful shorthand.

¹³Later work would show that no effect existed.

Fig. 7.3 $\log_{10} \alpha$ vs $\log_{10}(\lambda/1m)$. α , the strength of the Fifth Force, is constrained to lie below the curves. λ is the range of the force. From Gibbons and Whiting (1981).



7.1.3 The Fifth Force

Until early 1983 the two strands, that of the energy dependence of the CP-violating parameters in K -meson decay and that of modifications of Newtonian gravity and their experimental tests, proceeded independently. At about this time, Fischbach became aware of the discrepancies between the laboratory and geophysical measurements of G and the anomalies for gravitational theory. He made no connection between the two problems because he was still thinking in terms of a long-range force, which had been experimentally ruled out for CP violation. In early 1984 he realized that this would not apply to a short-range force and that the effect could be much smaller. At this time he also became aware of the summary by Gibbons and Whiting, which did not rule out such a force. He realized that a short-range force might be a common solution to both problems.

Fischbach, Aronson, and their collaborators looked for other places in which such an effect might be seen with existing experimental sensitivity. They found only three:

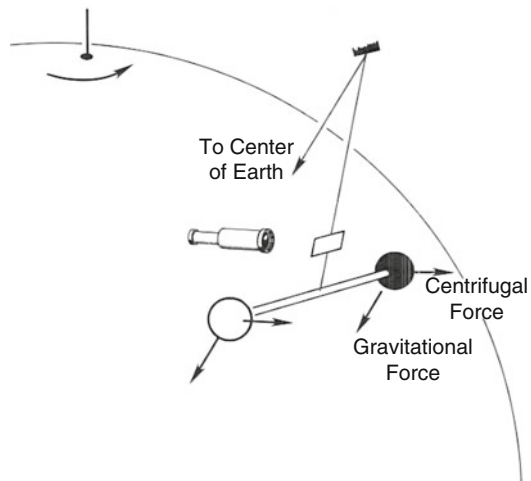
- 1) the K^0 -meson system at high energy, which they had already examined;
- 2) the comparison between satellite and terrestrial determinations of g , the local gravitational acceleration;

- 3) the original Eötvös experiment, which had measured the difference between the gravitational and inertial masses of different substances, and a set of an upper limit of five parts in 10^9 for that difference. If there were a short-range, composition-dependent force, then it might show up in this experiment.

The apparent energy dependence of the CP-violating parameters along with the discrepancy between gravitational theory and the mineshaft experiments led Fischbach and his colleagues to reexamine the original data of Eötvös et al. (1922) to see if there was any evidence for a short-range, composition-dependent force. By this time they knew of Holding's and Tuck's result which gave G measured in a mine as $G = (6.730 \pm 0.003) \times 10^{-11} \text{ m}^3 \text{ kg}^{-1} \text{ s}^{-2}$ in disagreement with the best laboratory value of $(6.6726 \pm 0.0005) \times 10^{-11} \text{ m}^3 \text{ kg}^{-1} \text{ s}^{-2}$. This result was, however, still uncertain because of possible regional gravity anomalies. Fischbach and colleagues used a modified gravitational potential $V = -GmM/r[1 + \alpha e^{-r/\lambda}]$, which they remarked could explain the geophysical data if $\alpha = (-7.2 \pm 3.6) \times 10^{-3}$ and $\lambda = 200 \pm 50 \text{ m}$. This was from a private communication from Stacey. Details appeared later in Holding et al. (1986). This result was within the window found by Gibbons and Whiting. This potential had the same mathematical form as that suggested much earlier by Fujii. Recall that Fujii had also suggested redoing the Eötvös experiment.¹⁴

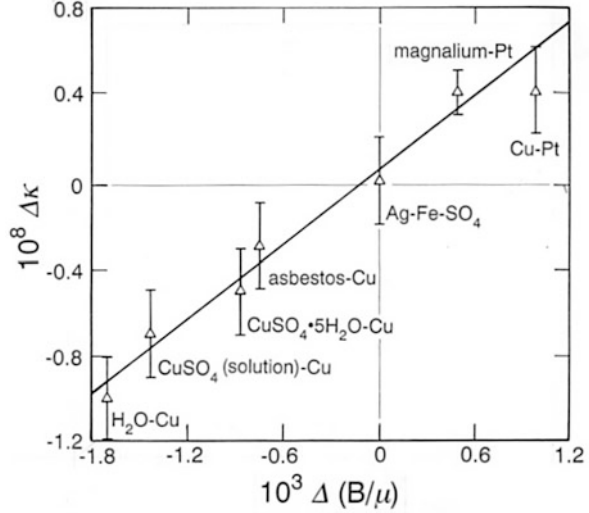
The apparatus for the Eötvös experiment is shown schematically in Figure 7.4. One can see that because of the rotation of the Earth, the gravitational force is not parallel to the fiber. If the gravitational force on one of the masses differs from

Fig. 7.4 A schematic view of the Eötvös experiment. From Will (1984).



¹⁴Fischbach has stated that Fujii's work had no direct influence on this work. He keeps detailed chronological notes of papers read. He reports that he has notes on Fujii's work at this time, but does not recall it having any influence on his work.

Fig. 7.5 $\Delta\kappa$, as a function of $\Delta(B/\mu)$. From Fischbach et al. (1986b).



that on the other mass or if the ratio of the gravitational to inertial mass of the two objects differs, then the rod will rotate about the fiber axis. Fischbach and colleagues attempted to find a single explanation for the gravitational discrepancies and the apparent energy dependence of the CP-violating parameters. They found that if they considered a hypercharge field with a small, finite mass hyperphoton (the K^0 and anti- K^0 have opposite hypercharges), they obtained a potential of the same mathematical form as shown above. They also found that $\Delta\kappa = \Delta a/g$, the fractional difference in gravitational acceleration for two substances, was proportional to $\Delta(B/\mu)$ for the two substances, where B was the baryon number and μ was the mass of the substance in units of the mass of atomic hydrogen.

Their reanalysis of the Eötvös data is shown in Figure 7.5 (Fischbach et al. 1986a).¹⁵ The clear linear dependence seen, showing a composition dependence, is supported by a least-squares fit to the equation $\Delta\kappa = a\Delta(B/\mu) + b$. They found $a = (5.65 \pm 0.71) \times 10^{-6}$ and $b = (4.83 \pm 6.44) \times 10^{-10}$. This is an eight-standard deviation difference from the zero expected from Newtonian gravity or general relativity, which are both composition independent. They concluded, “We find that the Eötvös-Pekar-Fekete data are sensitive to the composition of the material used,

¹⁵An interesting aspect of this reanalysis was reported in a footnote to this paper. Rather than reporting the observed values of $\Delta\kappa$ for the different substances directly, Eötvös and his colleagues had presented their results relative to platinum as a standard. “The effect of this combining say $\Delta\kappa(H_2O - Cu)$ and $\Delta\kappa(Cu - Pt)$ to infer $\Delta\kappa(H_2O - Pt)$ is to reduce the observed effect (for water and platinum) from 5σ to 2σ ” (Fischbach et al. 1986a). $\Delta\kappa(H_2O - Cu) = (-10 \pm 2) \times 10^{-9}$ and $\Delta\kappa(Cu - Pt) = (+4 \pm 2) \times 10^{-9}$, respectively. Adding them to obtain $\Delta\kappa(H_2O - Pt)$ yields $(-6 \pm 3) \times 10^{-9}$. Fischbach and colleagues chose to use copper as their standard which minimized the need for such additions.

and that their results support the existence of an intermediate-range coupling to baryon number or hypercharge” (Fischbach et al. 1986a, p. 3).¹⁶ They calculated the coupling constant for their new interaction for both the Eötvös data and for the geophysical data and found that they differed by a factor of 15, which they found “surprisingly good” in view of the simple model of the Earth they had assumed. As discussed below, not everyone was so sanguine about this.

It seems fair to summarize the paper of Fischbach and his colleagues as follows. A reanalysis of the original Eötvös paper presented a suggestive evidence for an intermediate-range, composition-dependent force. With a suitable choice of parameters (a force about one percent of the gravitational force with a range of approximately 100 meters), they could relate this force to measurements of gravity in mineshafts and to a suggested energy dependence in the parameters of the neutral K -meson system.

7.2 ...and Fall

7.2.1 *The Immediate Reaction*

The suggestion by Fischbach and his colleagues had an immediate impact in the popular press. On January 8, 1986, only 2 days after the publication of their paper, a headline in the *New York Times* announced, “Hints of Fifth Force in Nature Challenge Galileo’s Findings.” This referred to the composition dependence of the suggested force, which implied that different substances would fall at different rates. This would disagree with what Galileo was supposed to have observed at the Leaning Tower of Pisa.¹⁷ This was the naming of the “Fifth Force.” On January 15 an editorial in the *Los Angeles Times* also discussed the subject. It cited the skepticism of Richard Feynman, a Nobel Prize winner in physics. Feynman’s skepticism concerned the factor of 15 difference (a more careful analysis gave a factor of 30) between the force needed to explain the Eötvös data and that needed to explain the gravitational mine data. Feynman argued that the geophysical results already showed that the hypothesis was incorrect.

The battle would not, however, be conducted or decided in the popular press but rather in the technical literature. One of the most important early developments was the recognition that local mass asymmetries, such as cliffs, hills, or large buildings, were of crucial importance not only in the reanalysis of the Eötvös experiment but

¹⁶A skeptic might remark that the effect is seen only when the data are plotted as a function of $\Delta(B/\mu)$, a theoretically suggested quantity. As Alvaro De Rujula remarked, “In that case, Eötvös and collaborators would have carried their secret to their graves: how to gather ponderous evidence from something like baryon number decades before the neutron was discovered” (De Rujula 1986a, p. 761). Although one may be surprised, along with De Rujula, that data taken for one purpose takes on new significance in the light of later experimental and theoretical work, it is not unprecedented.

¹⁷There is some question as to whether Galileo ever performed this experiment. See Cooper (1935).

also in the design of experiments to search for the Fifth Force. Thodberg (1986) pointed out that the Eötvös reanalysis required an attractive Fifth Force, whereas the geophysical results required a repulsive Fifth Force. Fischbach and colleagues remarked that Thodberg was indeed correct but that further analysis had shown “that one cannot in fact deduce from the EPF [Eötvös-Pekar-Fekete] data whether the force is attractive or repulsive. The reason for this is that in the presence of an intermediate-range force, local horizontal mass inhomogeneities (e.g., buildings or mountains) can be the dominant source in the Eötvös experiment” (Fischbach et al. 1986b, p. 2464). In order to determine the magnitude and sign of the effect, one needed more detailed knowledge of the local mass distribution than was then available. Fischbach and his collaborators even searched for a detailed map of the University of Budapest campus, where Eötvös had done his work. They also tried to discover whether the building in which the experiment was done had a basement, which would influence the local mass distribution. The importance of the local mass distribution could also explain the numerical discrepancy between the force derived from the Eötvös reanalysis and that found from the mine data that had bothered Feynman and others.

Other authors suggested redoing the Eötvös experiment by placing the torsion balance on a high cliff or in a tunnel in such a cliff (Bizzeti 1986; Milgrom 1986; Neufeld 1986; Thieberger 1986; De Rujula 1986a,b). They claimed that such a location, which had a large local mass asymmetry, could increase the sensitivity of the experiment by a factor of 500. De Rujula (1986a) and Eckhardt (1986) argued that the original Eötvös reanalysis would not have been at all sensitive to a Fifth Force without local mass inhomogeneities. They noted that for a deformed rotating Earth, the fiber is perpendicular to the deformed surface. For a homogeneous Earth, the symmetry of the local matter distribution will give no net force on the balance. De Rujula quipped, “Although malicious rumor has it that Eötvös himself weighed more than 300 pounds [suggesting that Eötvös himself was the source of a local mass asymmetry], unspecific hypotheses are not, a priori, particularly appealing” (De Rujula 1986a, p. 741). De Rujula’s quip is completely without merit. Eötvös was a mountain climber, and photographs indicate rather clearly that he did not weigh 300 pounds. In fact, a peak in the Dolomites is named for him.

The initial reanalysis of the Eötvös experiment was incorrect because it did not consider local mass asymmetries. The subsequent criticism not only modified the theoretical model but also allowed one to design experiments that would be far more sensitive to the presence of the hypothesized Fifth Force. Other critics suggested that there was, in fact, no observed effect and that Fischbach and his colleagues had made an error in the reanalysis. De Rujula, however, performed his own reanalysis of the Eötvös data and obtained results identical to those of Fischbach and collaborators.¹⁸

¹⁸De Rujula’s analysis was important because it answered the question of whether one should use reduced mass. In several measurements Eötvös used a brass vial to hold the sample of the material. In reporting the final results, he multiplied the measured value $\Delta\kappa$ by a factor $(M_{Sample} + M_{Container})/M_{Sample}$. This assumed that the container had no effect on the measurement. This was a reasonable procedure if one was interested only in setting an upper limit but might overestimate

Some physicists suggested that experiments on K mesons had already ruled out the Fifth Force. Questions were also raised as to whether one could explain the Eötvös results in terms of more conventional physics, without invoking a new force. (For details see Franklin 1993.)

Although the criticism may have made the reanalysis of the Eötvös data somewhat uncertain, it did not prevent physicists from planning new, more sensitive versions of old experiments and designing new ones to test for the presence of the Fifth Force. At the same time, theoretical physicists were attempting to find an explanation for the force and to see if it had implications in other areas. Unfortunately in all of these theoretical studies, the expected effects were quite small and did not suggest new experimental tests.

At the end of 1986, the evidential context for the Fifth Force was much the same as it had been on January 6, 1986, when Fischbach and colleagues had first published it. By early 1986 the inverse-square law of gravity had been tested at very short distances and had been confirmed, but the possibility of an intermediate-range force remained. Doubts had been raised about the proposed mechanism of the force, but other explanations were possible. The tantalizing effects of the reanalysis of the Eötvös experiment, the K -meson parameters, and the measurements of gravity in mineshafts still remained.

The attitude of scientists toward the Fifth Force at this time varied from outright rejection to regarding it as highly suggestive and plausible. Sheldon Glashow, a Nobel Prize-winning theoretical physicist, was quite negative. “Unconvincing and unconfirmed kaon data, a reanalysis of the Eötvös experiment depending on the contents of the Baron’s wine cellar [an allusion to the importance of local mass inhomogeneities], and a two-standard-deviation geophysical anomaly! Fischbach and his friends offer a silk purse made out of three sows ears, and I’ll not buy it” (quoted in Schwarzschild (1986, p. 20)). John Maddox noted that, “Fischbach et al. have provided an incentive for the design of better measurements by showing what kind of irregularity it will be sensible to look for” (1986, p. 173). An important feature of experimental design is knowing how large the observed effect is supposed to be. A much more positive view was, “Considerable, and justified, excitement has been provoked by the recent announcement—that a reanalysis of the Eötvös experiment together with recent geophysical gravitational measurements supports the existence of a new fundamental interaction” (Lusignoli and Pugliese 1986, p. 468).

It seems clear, judging by the substantial amount of work published in 1986, that a significant segment of the physics community thought the Fifth Force hypothesis was plausible enough to be worthy of further investigation. Although almost invisible in the published literature, experiments were being designed, performed, and analyzed. The results would start to appear in early 1987.

the effect. Fischbach and collaborators had used the “composite” value, whereas De Rujula used the reduced value (vials not included). The agreement of the two slopes showed that the analysis was independent of which one used, as long as one remained consistent.

7.2.2 A Composition-Dependent Force? Was Galileo Wrong?

There would be two sets of discordant experimental results that had to be resolved in order to decide whether there was a Fifth Force. The first strand of experimental investigation of the Fifth Force was the search for a composition dependence of the gravitational force. (The second involved the question of a proposed deviation from Newton's inverse-square law of gravity and is discussed below.) The former were the first published experimental results. Recall that the strongest piece of evidence cited when the Fifth Force was initially proposed came from a reanalysis of the Eötvös experiment. That reanalysis had shown a large and surprising composition-dependent effect. This was the effect that was subsequently investigated. Both types of experiment are shown in Figure 7.6.

Two types of composition-dependence experiments are shown in the top row. In order to observe the effect of a short-range force such as the Fifth Force, one needs a local mass asymmetry. This asymmetry was provided by either a terrestrial source—a hillside or a cliff—or by a large, local, laboratory mass. If there were a composition-dependent, short-range force, the torsion pendulum made of two different substances would twist. A variant of this experiment was the float experiment, in which an object floated in a fluid and in which the difference in gravitational force on the float and on the fluid would be detected by the motion of the float. These were done with terrestrial sources.

The results of the first tests for a composition-dependent force appeared in January, 1987, 1 year after the Fifth Force first appeared in print. They disagreed. Peter Thieberger, using a float experiment, found results consistent with the presence

Fig. 7.6 Different types of experiment to measure the Fifth Force. The upper row shows composition-dependence experiments. The bottom row shows distance-dependence experiments. The left column shows terrestrial sources; the right column shows laboratory/controlled sources. From Stubbs (1990).

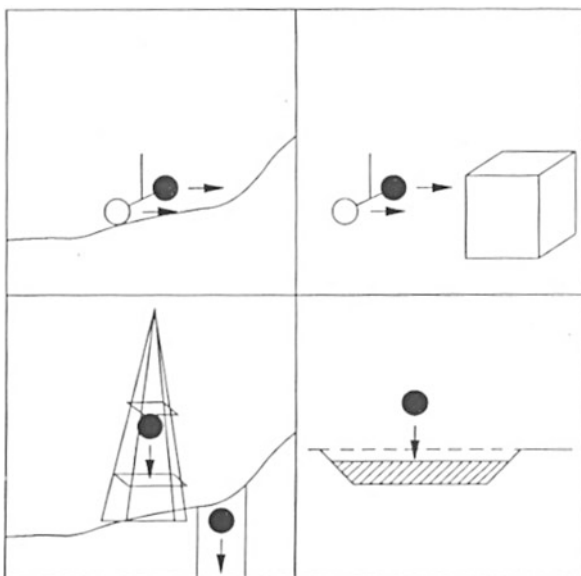
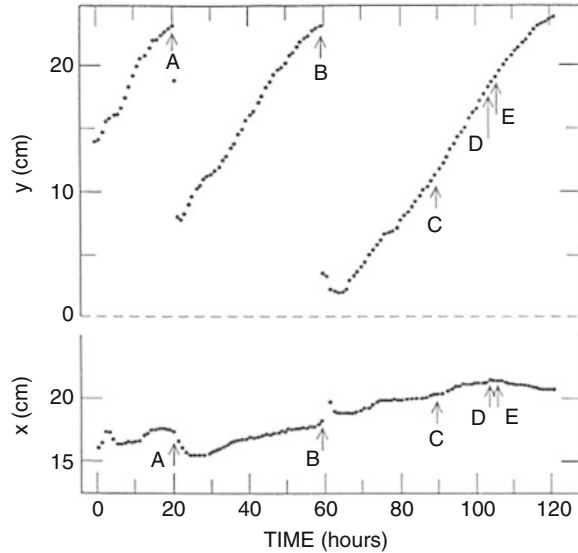


Fig. 7.7 The position of the center of the copper sphere as a function of time. The y axis points away from the cliff. The position of the sphere was reset at points A and B . From Thieberger (1987).

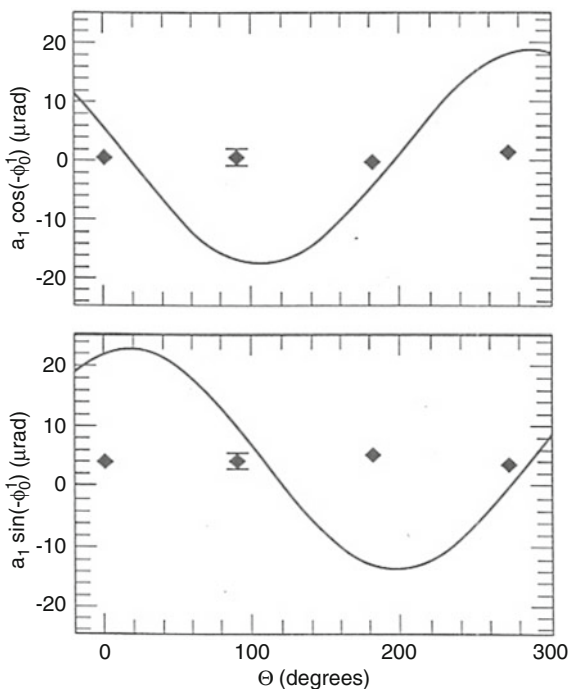


of such a force (Thieberger 1987). A group at the University of Washington, headed by Eric Adelberger and whimsically named the Eöt-Wash group, found no evidence for such a force and set rather stringent limits on its presence (Adelberger et al. 1987).

The results of Thieberger's experiment, performed on the Palisades cliff in New Jersey, are shown in Figure 7.7. Thieberger measured the difference in force on the copper float and on the water. One can see that the float moves quite consistently and steadily away from the cliff (the y -direction) as one would expect if there were a Fifth Force. (One wag remarked that all the experiment showed was that any sensible float wanted to leave New Jersey.) Thieberger eliminated other possible causes for the observed motions. These included the possible effects of magnetic forces, thermal gradients, and leveling errors. No significant effects were observed. He also rotated his apparatus by 90° to check for possible instrumental asymmetries and obtained the same positive result. In addition, he performed the same experiment at another location, one without a local mass asymmetry or cliff, and found no effect, as expected. He concluded, "The present results are compatible with the existence of a medium-range, substance-dependent force which is more repulsive (or less attractive) for Cu than for H_2O Much work remains before the existence of a new substance-dependent force is conclusively demonstrated and its properties fully characterized" Thieberger (1987, p. 1068).

The Eöt-Wash experiment used a torsion pendulum located on the side of a hill on the University of Washington campus. If the hill attracted the copper and beryllium test bodies, used in the apparatus, differently, then the torsion balance would experience a net torque. None was observed (Figure 7.8). The group also eliminated other possible causes of effects that might either mimic the presence of a

Fig. 7.8 Deflection signal as a function of Θ . The theoretical curves correspond to a Fifth Force with a strength $\alpha = 0.01$ and a range $\lambda = 100\text{ m}$. From Raab (1987).



Fifth Force or mask the effects of such a force. The possible effects of electrostatic forces, instrumental asymmetries, magnetic forces, gravity gradients, and the tilt of the apparatus were measured and shown to be negligible.

The discordant results were an obvious problem for the physics community. Both experiments appeared to have been carefully done, with all plausible and significant sources of possible error and background adequately accounted for. Yet the two experiments disagreed. In this case we are dealing with attempts to observe and measure the same quantity, a composition-dependent force, with very different apparatuses, a float experiment, and a torsion pendulum. Was there some unknown but crucial background in one of the experiments that produced the wrong result? To this day, no one has found an error in Thieberger's experiment, but the consensus is that the Eöt-Wash group is correct and that Thieberger is wrong—that there is no Fifth Force. How was the discord resolved?

In this episode it was resolved by an overwhelming preponderance of evidence. The torsion pendulum experiments were repeated by others including Fitch et al. (1988), Cowsik et al. (1988), Bennett (1989), Boynton (1990), Boynton et al. (1987), Boynton and Peters (1989),¹⁹ and Newman (Newman et al. 1989; Nelson et al.

¹⁹Boynton had initially found a 3.5 standard-deviation positive effect. His later, more accurate experiments found no effect.

1990), and by the Eöt-Wash group (Adelberger 1988, 1989; Heckel et al. 1989; Stubbs et al. 1989). None gave evidence for a Fifth Force.

Bennet’s experiment is particularly interesting. He reported a measurement of the difference in force exerted on copper and lead masses by a known mass of water, located nearby. The experiment used a torsion balance located near the Little Goose Lock on the Snake River in eastern Washington, in which the water level was changed periodically to allow the passage of boats. This change in water level provided the known mass of water. The difficulty of real, as opposed to ideal, experiments is clearly illustrated in this experiment. “Because the data were taken during a dry period (August 1988), separate lock fillings could not be made just for the experiment. On average there were four lockages a day from barge traffic which could occur at any hour of the day or night with only a half-hour advance notice.” The apparatus needed minor adjustment every 4 or 5 hours and then took about 2 hours to stabilize, allowing good data to be taken for the next 2 or 3 hours. “The success of a particular run depended on the coincidence of this observation period with the arrival of lock traffic and, typically only one could be observed in a period of about 6 h during weekdays. Fortunately, traffic on weekends was heavier because of pleasure craft. Although consistent with individual isolated experiments, by far the best data were obtained on Sunday, 21 August 1988, when an armada of small craft went up and down the river” (Bennett 1989, p. 367).

All of the repetitions, in different locations and with different substances, gave consistently negative results. There was also evidence against the Fifth Force from modern versions of Galileo’s Leaning Tower of Pisa experiment performed by Kuroda and Mio (1989a,b, 1990) and by Faller and his collaborators (Niebauer et al. 1987; Speake et al. 1990). As more negative evidence was provided, the initial and startling effect claimed by Fischbach and collaborators became less and less dramatic (Figures 7.9 and 7.10). In fact, one might reasonably say that the effect had disappeared. In addition, Bizzeti, using a float apparatus similar to that used by

Fig. 7.9 Comparison of the Eötvös reanalysis of Fischbach et al. with the results of the Eöt-Wash I and III experiments. The error bar on the Eöt-Wash III datum is smaller than the dot. From Adelberger (1989).

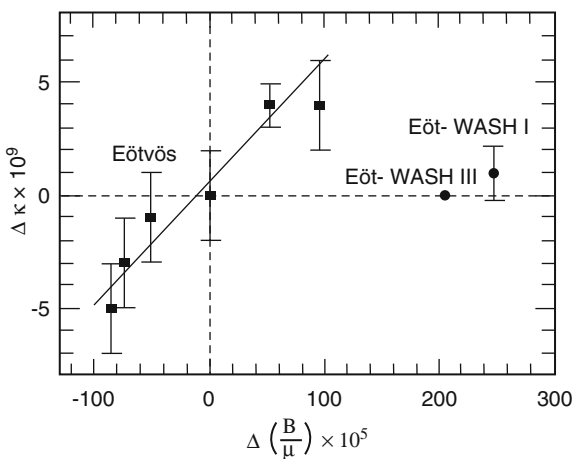


Fig. 7.10 The results of Kuroda and Mio added to Figure 7.9.

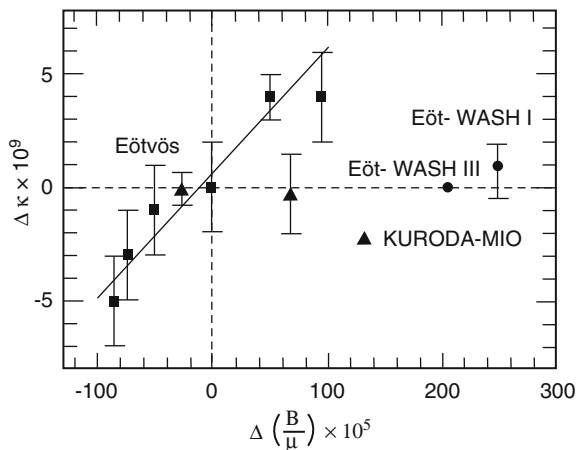
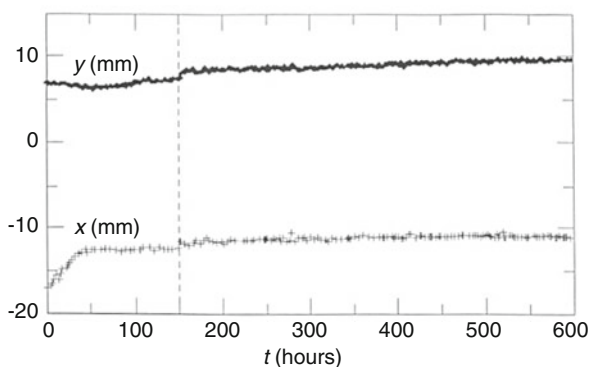


Fig. 7.11 The position of the sphere completely immersed in liquid as a function of time. The vertical line marks the time at which the restraining wires were removed. From Bizzeti et al. (1988).



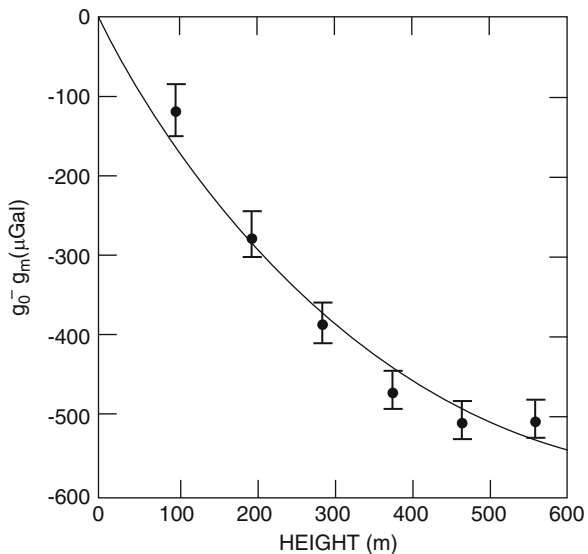
Thieberger, also obtained results showing no evidence of a Fifth Force (Bizzeti et al. 1988, 1989a,b). (Compare Bizzeti's results (Figure 7.11) with those of Thieberger (Figure 7.7)). Bizzeti's result was quite important. Had he agreed with Thieberger, then one might well have wondered whether there was some systematic difference between torsion balance experiments and float experiments that gave rise to the conflicting results. This did not happen. There was an overwhelming preponderance of evidence against composition dependence of the Fifth Force. Even Thieberger, although he had not found any error in his own experiment, agreed. "Unanticipated spurious effects can easily appear when a new method is used for the first time to detect a weak signal. . . Even though the sites and the substances vary, effects of the magnitude expected [from his initial experiment] have not been observed. . . It now seems likely that some other spurious effect may have caused the motion observed at the Palisades cliff" (Thieberger 1989, p. 810).

7.2.3 Towers and Mineshafts: The Distance Dependence of the Gravitational Force

A second way in which the presence of the Fifth Force could be tested was by investigating the distance dependence of the gravitational force, to see if there was a deviation from Newton's inverse-square law. This type of experiment measured the variation of gravity with position, usually in a tower, in a mineshaft, or in a borehole (Figure 7.6, bottom row). All of the experiments used a standard device, a LaCoste-Romberg gravimeter, to measure gravity. The measurements were then compared with the values calculated using a model of the Earth, surface gravity measurements, and Newton's law of gravitation. This type of calculation had been done often and was regarded as reliable. The results of the calculation were, however, quite sensitive to the surface gravity measurements and to the model of the Earth used. This made knowledge of the local mass distribution and of the local terrain very important.

Evidence from such measurements had provided some of the initial support for the existence of the Fifth Force. Geophysical measurements during the 1970s and 1980s had given values of G , the universal gravitational constant, that were consistently higher, by about 1%, than that obtained in the laboratory. Because of possible local mass anomalies, they were also "tantalizingly uncertain." After the proposal of the Fifth Force, further experimental work was done. At the Moriond Workshop in January, 1988, Donald Eckhardt presented results from the first of the new tower gravity experiments (Eckhardt et al. 1988, 1989). The results differed from the predictions of the inverse-square law by $-500 \pm 35 \mu\text{Gal}$, ($1 \mu\text{Gal} = 10^{-8} \text{ms}^{-2}$) at the top of the tower (Figure 7.12).

Fig. 7.12 Eckhardt's results for the difference between the measured and calculated values of g , the acceleration due to gravity, as a function of height. From Fairbank (1988).



Further evidence for the Fifth Force was provided by a group that measured the variations in gravity in a borehole in the Greenland ice cap (Ander et al. 1989). They found an unexplained 3.87 mGal discrepancy between the measurements taken at a depth of 213 m and those taken at a depth of 1673 m . This was larger and opposite in sign to the geophysics results of Stacey and collaborators. The experimental advantage of the Greenland experiment was the uniform density of the ice cap. The disadvantages were the paucity of surface gravity measurements and, as the group noted, the presence of underground geological features that could produce gravitational anomalies.

The Livermore group, using measurements taken at the BREN Tower at the Nevada test site, found a 2.5% discrepancy between the observed gravity gradient and that predicted by a standard Newtonian model of the Earth (Thomas et al. 1988). This result disagreed in magnitude with Stacey's 0.52% discrepancy and, in both sign and magnitude, with Eckhardt's 0.29% discrepancy. They concluded, however, "that the model [of the Earth] does not reflect the total mass distribution of the Earth with sufficient accuracy to make a statement about Newtonian gravity [or about the Fifth Force]" (Thomas et al. 1988, p. 591). The evidence from tower and mineshaft experiments prior to 1988 was consistent with the Fifth Force, albeit with considerable uncertainty. There was, however considerable, although not unambiguous negative evidence from other types of experiment. Negative evidence from tower experiments would, however, be forthcoming, and it is the discrepancy between the tower results that I will address here.

Even before those negative results appeared, questions and doubts were raised concerning the positive results. It was not, in fact, the gravity measurements themselves that were questioned. These were all obtained with a standard and reliable instrument. It was, rather, the theoretical calculations used for the theory-experiment comparison that were criticized. One of the important features needed in these calculations was an adequate model of the Earth.

The Greenland group's calculation was the first to be criticized. It was subjected to severe criticism, particularly for the paucity of surface gravity measurements near the location of their experiment (their survey included only 16 such points) and for the inadequacy of their model of the Earth. It was pointed out that there were underground features in Greenland of the type that could produce such gravitational anomalies. The group later admitted that their result could be interpreted either as evidence for non-Newtonian gravity (a Fifth Force) or explained by local density variations. "We cannot unambiguously attribute it to a breakdown of Newtonian gravity because we have shown that it might be due to unexpected geological features below the ice" (Ander et al. 1989, p. 985).

Robert Parker, a member of the Greenland group, as well as David Bartlett and Wesley Tew, suggested that both the positive evidence for the Fifth Force of Eckhardt and collaborators and that from the mineshaft experiments could be explained by either local density variations or by inadequate modeling of the local terrain (Parker and Zumberge 1989; Bartlett and Tew 1989a).

Eckhardt disagreed. His group presented a revised, and lower, value for the deviation from Newtonian gravity at the top of their tower of $350 \pm 110\mu\text{Gal}$

(Eckhardt et al. 1989). They attributed this change, a reduction of approximately one-third, to better surface gravity data and to finding a systematic elevation bias in their previous survey. (Gravity measurements tend to be made on roads rather than in ditches or surrounding fields. Roads are usually higher than their surroundings, giving rise to an elevation bias.) “We also had the help of critics who found our claims outrageous.” They concluded that, “nevertheless the experiment and its reanalysis are incomplete and we are not prepared to offer a final result” (Eckhardt et al. 1989, p. 526).

The Lawrence Livermore Laboratory group presented a result from their gravity measurements at the BREN tower at the Nevada test site (Kasameyer et al. 1989). To overcome the difficulties with their previous calculations, they had extended their gravity survey to include 91 of their own gravity measurements within 2.5 km of the tower, supplemented with 60000 surface gravity measurements within 300 km, done by others. Contrast this with the 16 points in the Greenland survey. They presented preliminary results in agreement with Newtonian gravity, reporting that, at the top of the tower, there was no difference between the measured and predicted values.

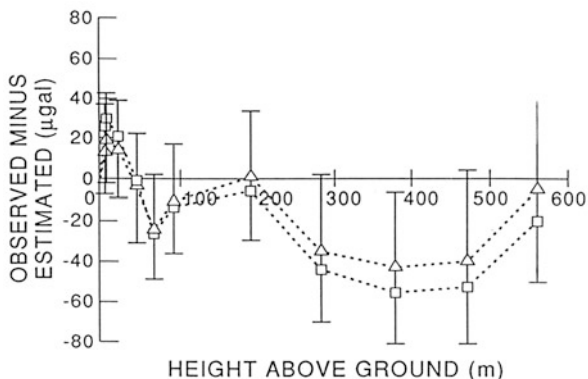
Bartlett and Tew (1989b, 1990) continued their work on the effects of local terrain. They argued that the Hilton mine results of Stacey and his collaborators could also be due to a failure to include local terrain in their theoretical model. They communicated their concerns to Stacey privately. Their view was confirmed when, at the General Relativity and Gravitation Conference in July 1989, G. J. Tuck reported that their group had incorporated a new and more extensive surface gravity survey into their calculation. Preliminary analysis of these data indicated a regional bias that reduced the anomalous gravity gradient to two-thirds of the value that they had previously reported (with a 50% uncertainty). With such a large uncertainty, the results of Stacey and his collaborators could no longer be considered as support for the Fifth Force.

Parker and Mark Zumberge, two members of the Greenland group, offered a general criticism of tower experiments. They argued, in some detail, that they could explain the anomalies reported in both Eckhardt’s tower experiment and in their own ice cap experiment, using conventional physics and plausible local density variations. They concluded that there was “no compelling evidence for non-Newtonian long-range forces in the three most widely cited geophysical experiments [those of Eckhardt, of Stacey, and their own]. . . and that the case for the failure of Newton’s Law could not be established” (Parker and Zumberge 1989, p. 31).

The last hurrah for tower gravity experiments that supported the Fifth Force was signaled in the paper, “Tower Gravity Experiment: No Evidence for Non-Newtonian Gravity” (Jekeli et al. 1990). In this paper Eckhardt’s group presented their final analysis of their data, which included a revised theoretical model, and concluded that there was, in fact, no deviation from Newtonian gravity. (See Figure 7.13, and contrast this with their initial positive result shown in Figure 7.12). Two subsequent tower results also supported Newton’s Law.

The discord had been resolved. The tower and mineshaft measurements were correct. It was the comparison between theory and experiment that had led to the

Fig. 7.13 Difference between measured and calculated values of g as a function of height. From Jekeli et al. (1990).



discord. It had been shown that the results supporting the Fifth Force could be explained by inadequate theoretical models, either failure to account adequately for local terrain or the failure to include plausible local density variations.

Scientists make decisions in an evidential context. The Fifth Force was a modification of Newtonian gravity. Newtonian gravity and its successor, general relativity, were strongly supported by other existing evidence. In addition, there were other credible negative tower gravity results that did not suffer from the same difficulties as did the positive results. There was also, as discussed earlier, an overwhelming preponderance of evidence against the Fifth Force from other types of experiment. The decision as to which theory-experiment comparison was correct was not made solely on the basis of the experiments and calculations themselves, although one could have justified this. Scientists examined all of the available evidence and came to a reasoned decision about which were the correct results—and concluded that the Fifth Force did not exist.

In both instances discussed in this paper, the composition dependence and the distance dependence of the proposed Fifth Force, the decision that such a force did not exist was made on the basis of reasons that allow us to consider experimental results as the basis for scientific knowledge. In the case of the distance dependence, it was shown that the positive results had overlooked effects in their theoretical calculations that resulted in an incorrect experiment-theory comparison. This, combined with credible negative results, argued against the existence of the Fifth Force. The discrepancy between the Thieberger and Adelberger results on the composition dependence of the Fifth Force was resolved by an overwhelming preponderance of evidence. In addition, Bizzeti and collaborators, using an apparatus quite similar to that of Thieberger, found no evidence for the Fifth Force. This argued against any crucial difference between the different types of apparatus being responsible for the discordant results.

In 1990, at a Moriond Workshop attended by most of those working in the field, Orrin Fackler of the Livermore group remarked, “The Fifth Force is dead.” No one disagreed. The Fifth Force is not with us.²⁰

7.3 Epilogue: The Fifth Force Since 1991

We left our story at the 1990 Moriond Workshop with the stated demise of the Fifth Force. As even Ephraim Fischbach and Carrick Talmadge, two of the proposers of the initial hypothesis remarked, “No compelling evidence has yet emerged that would indicate the presence of a fifth force, . . .” (Fischbach and Talmadge 1992, p. 214).

Despite these obituaries, work on the Fifth Force, both experimental and theoretical, has continued into the twenty-first century. This includes explicit tests of the hypothesis. Other works, on the universality of free fall, on possible violation of Newton’s inverse-square law of gravity and on the weak equivalence principle in general relativity also have relevance for the Fifth Force. These later papers, although relevant, do not always mention the Fifth Force explicitly or cite the initial paper of Fischbach and his collaborators. Thus, the Eöt-Wash collaboration stated,

The universality of free fall (UFF) asserts that a point test body, shielded from all known interactions except gravity, has an acceleration that depends only on its location. The UFF is closely related to the gravitational equivalence principle, which requires an exact equality between gravitational mass m_g and inertial mass m_i and therefore the universality of gravitational acceleration. Experimental tests of the UFF have two aspects —they can be viewed as tests of the equivalence principle or as probes for new interactions that violate the UFF. (Su et al. 1994, p. 3614)

The UFF test would also test for the Fifth Force. The paper of Su et al. quoted above, for example, sets limits on possible violations of Newton’s law of universal gravitation and on a possible Fifth Force, but did not cite the 1986 paper of Fischbach and collaborators.

In this epilogue I will concentrate on the experimental work that has relevance for the Fifth Force which has taken place since the funeral at Moriond. This is not intended to be a complete history but rather to give the flavor of the variety of experimental work done on the Fifth Force at the end of the twentieth century and the beginning of the twenty-first century. We will find that the Fifth Force is still dead.²¹

²⁰With apologies to George Lucas.

²¹I will not discuss several fascinating proposed experiments, which were never performed. For details of these proposals and for a more detailed history, see (Franklin and Fischbach 2016).

7.3.1 The 1990s

One of the earliest of these later experiments was performed by a group in China (Yang et al. 1991). The experimenters measured the differences in the acceleration due to gravity at various distances from an empty oil reservoir caused by filling or emptying the reservoir with water.²² The acceleration was measured with a LaCoste-Romberg gravimeter, the standard apparatus used in earlier tower experiments. The experimenters compared the measured differences in acceleration with those calculated from Newtonian gravity alone. Any difference would be attributed to the Fifth Force. Their results are shown in Table 7.1. No differences between the measured and calculated values were seen. The group concluded, “It is worth pointing out that a weak intermediate-range interaction of Yukawa form is not excluded by our data but the possible strength of such an interaction is highly constrained $|\alpha| < 0.002$. This is in agreement with the results of the WTVD [Eckhardt’s group] and BREN [Lawrence Livermore group] tower gravity experiments” (Yang et al. 1991, p. 332).²³

There were also replications of previous types of experiment. Liu et al. (1992) measured the acceleration due to gravity as a function of height on a 320 m tower. This would test the possible distance dependence of the Fifth Force. They noted the previous discord between the early positive results reported by Eckhardt and his collaborators and the negative results reported by the Lawrence Livermore group, by Speake et al. (1990), by the later results of Eckhardt’s group, and by others including Cruz et al. (1991). They remarked, “Many have questioned the results of Eckhardt et al. including Thomas et al. [the Livermore group] who, in an independent tower (BREN tower) experiment, found no evidence for non-Newtonian gravity. More recently Eckhardt et al. have revised their analysis and now their results appear consistent with Newtonian gravity. The newer and more precise Erie tower results

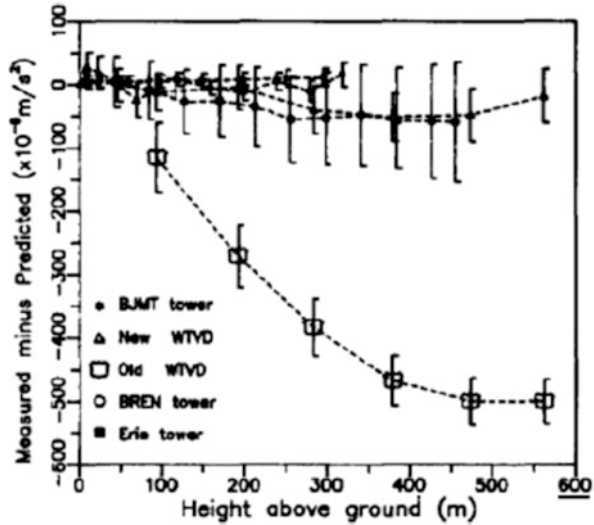
Table 7.1 Gravimetric measurements from Yang et al. (1991)

Distance from central axis of water cylinder (m)	Mean experimental value Δg_e and its standard deviation (10^{-5}m/s^2)	Newtonian prediction Δg_N (10^{-5}m/s^2)	$\Delta g_e / \Delta g_N$
10.00	0.424±0.002	0.423	1.002±0.005
20.00	0.273±0.002	0.272	1.004±0.007
30.00	0.146±0.002	0.145	1.007±0.014
40.00	0.075±0.002	0.073	1.027±0.027
50.00	0.040±0.003	0.038	1.053±0.079

²²This was similar to Bennett’s experiment at the lock on the Snake River, discussed earlier.

²³These were discussed earlier.

Fig. 7.14 Measured minus predicted values of the acceleration due to gravity as a function of the height aboveground for various tower experiments. From Liu et al. (1992).



of Cruz et al. (1991) now set a little stronger constraints on such a kind of non-Newtonian force. We decided that an independent experiment would help clarify the situation, and undertook to perform a tower test of gravity” (Liu et al. 1992, p. 131).

The experimenters used the standard LaCoste-Romberg gravimeter and corrected their results for tides, drift, gravimeter screw errors, and systematic effects due to tower motion. (All measurements were done at wind speeds less than 3 m/s). They stated that their tower was stable and located on a nearly flat terrain. Their results are shown in Figure 7.14 along with both the old and new results of Eckhardt et al. (1988) and several of the newer results. They concluded, “In a tower test of Newton’s inverse square law of gravitation we found no evidence for the non-Newtonian force, and the accuracy of the experiments constrains the Yukawa potential coupling constant $|\alpha|$ to be less than 0.0005” (Liu et al. 1992, p. 131).

Carusotto and et al. (1992) performed an interesting variant on the Galileo-type free fall experiments discussed earlier. They measured the angular acceleration of a disk which had a half-disk of aluminum and a half-disk of copper (Figure 7.15): “If there is a difference Δg in the free-fall acceleration of aluminum and copper, then the disk assembly experiences a torque and, therefore there is an angular acceleration of the disk assembly . . .” (Carusotto and et al. 1992, p. 1723). The disk would rotate. The acceleration was measured using laser light reflected from corner reflectors placed on the disk. The experimenters checked the sensitivity of their apparatus and looked for possible systematic effects by first making measurements with a disk made only of aluminum. They found $\Delta g/g = (3.2 \pm 9.5) \times 10^{-10}$, consistent with zero, demonstrating that there were no large systematic effects. Using the half-copper half-aluminum disk, they found $\Delta g/g = (8.5 \pm 9.5) \times 10^{-10}$ and $\Delta g/g = (-4.8 \pm 11.2) \times 10^{-10}$ with the disk reversed. They combined the two

sets of measurements and set a limit of $\Delta g/g = (2.9 \pm 7.2) \times 10^{-10}$. “The result is compatible with zero (no g violation) and it is in quite good agreement with the one obtained by Kuroda and Mio for the same materials” (Carusotto and et al. 1992, p. 1725).

Experimental tests of the Fifth Force hypothesis continued in 1993. The group at the Tata Institute, using a torsion pendulum, sets more stringent limits on the possible coupling to isospin. Their 2σ limit for the strength was $-5.9 \times 10^{-5} \leq \alpha_I \leq 3.44 \times 10^{-5}$, “the best upper limit on α_I for all the experiments so far” (Unnikrishnan 1993, p. 408). Carusotto and collaborators reported further results on their falling-disk experiment (Table 7.2). In this experiment they used a copper-tungsten disk, rather than a copper-aluminum disk. They concluded, “There is no evidence for any g -universality violation, at the level of μGal , at least with the Galileo-type experiment performed so far” (Carusotto et al. 1993, p. 357).

In 1994 Eckhardt’s group²⁴ published results on measurements of the acceleration due to gravity as a function of the height of the measurement on a tower, using a tower different from the one they had used in their previous experiments (Romaides

Fig. 7.15 Schematic diagram of the Galileo-type experiment for a disk composed of two different metals. From Carusotto et al. (1993).

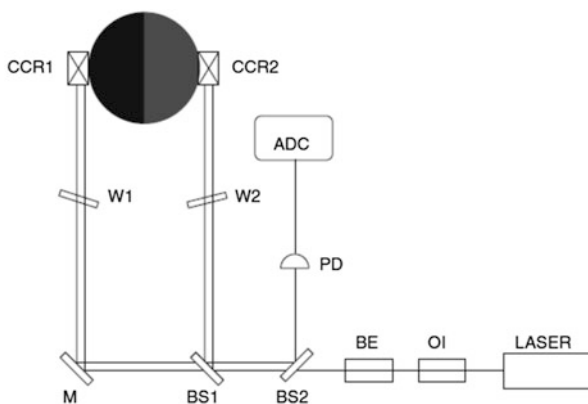


Table 7.2 From Carusotto et al. (1993)

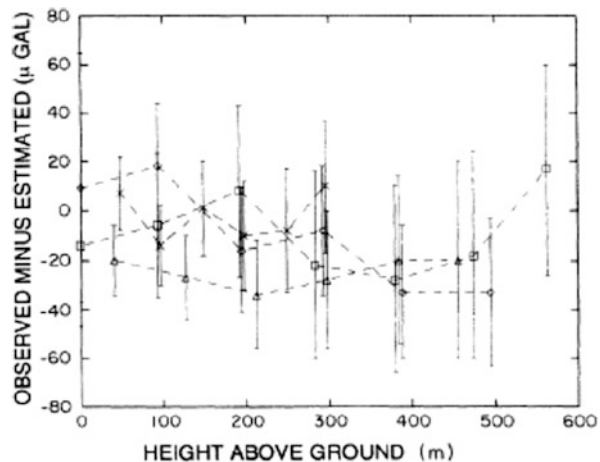
References	Compared materials	$\Delta g (\mu\text{Gal})$
Present work	Cu-W	0.71 ± 0.91
Carusotto and et al. (1992)	Al-Cu	0.29 ± 0.72
	Al-Cu	-0.13 ± 0.78
	Al-Be	0.43 ± 1.23
	Kuroda and Mio (1990)	Al-C
Niebauer et al. (1987)	Cu-U	0.13 ± 0.5

²⁴The group also included Fischbach and Talmadge, two of the initial proposers of the Fifth Force hypothesis.

et al. 1994). They noted that they had initially obtained results at the WTVD tower in North Carolina which showed an apparent violation of Newtonian gravitation but that their later results, along with those of other tower experiments, had shown that Newton's law of gravity was valid over a range from 10 m to 10 km. They stated that, "Two of the major difficulties in the experiment were the inaccessibility of some areas around the WTVD tower, and the lack of a good terrain model, which meant that some computations could not be done as rigorously as desired" (Romaides et al. 1994, p. 3608). Their new results were obtained at the WABG tower in Mississippi, which had the advantage of very flat local terrain and easy access for gravity measurements near the tower. They concluded, "The tower observations were compared to the predictions, with the largest discrepancy being $-33 \pm 30 \mu\text{Gal}$ at 493 m. The results are in good agreement with previous tower experiments, which also are in accord with the inverse-square law, and they set further restrictions on possible non-Newtonian forces" (Romaides et al. 1994, p. 3608). The group reported that their WABG results agreed not only with their last WTVD tower results but also with the results of other tower experiments (Figure 7.16). They stated that they were ending their investigations²⁵ and that "... we have learned from these and other experiments that there is no credible evidence for deviations from the inverse-square law over a laboratory to solar system scale length. By helping to fill in the scale $\lambda \approx 10^3$ m, tower experiments have thus played an important role in confirming our belief in the validity of Newtonian gravity" (Romaides et al. 1994, p. 3612).

The inclusion of tests of the Fifth Force as part of more general experimental work on general relativity and its implications became clear in the 1994 report of the Eöt-Wash group mentioned earlier (Su et al. 1994). The experimenters stated purpose was to measure the universality of free fall with respect to the Earth, the

Fig. 7.16 The observed-minus-model discrepancies for all tower experiments along with their associated 1 σ errors. The diamonds are the WABG results; the boxes are the WTVD results; the triangles are the BREN tower results; and the crosses are the Erie tower results. In order to avoid clutter, not all data points were plotted. Note the excellent agreement especially at the upper elevations. From Romaides et al. (1994).



²⁵As we shall see below, this is not quite accurate.

Table 7.3 Comparison of the 1991 (Adelberger et al. 1991) and 1994 (Su et al. 1994) Eöt-Wash results

	$\alpha \Delta(B/\mu)_{\text{detector}}$ $(B/\mu)_{\text{source}} \lambda = 30 \text{ m}$	$\lambda = 20 \text{ m}$	$\lambda = 50 \text{ m}$
1991	$(1.4 \pm 2.9) \times 10^{-8}$		
1991	$(-2.1 \pm 3.6) \times 10^{-8}$		
1994 (Be-Al detector)		$(-0.5 \pm 1.1) \times 10^{-8}$	$(-2.6 \pm 5.4) \times 10^{-9}$
1994 (Be-Cu detector)		$(-11 \pm 9.8) \times 10^{-9}$	$(-5.3 \pm 4.8) \times 10^{-9}$

Sun, our galaxy, and in the direction of the cosmic microwave dipole.²⁶ They further noted that, “Our galactic-source results tests the UFF [Universality of free fall] for ordinary matter attracted toward dark matter . . .” (Su et al. 1994, p. 3614).²⁷

The experimental group had made improvements in their torsion balance apparatus including better regulation of the turntable speed, compensation for gravity gradients, and in the calibration of their instruments. Although the Fifth Force is not explicitly mentioned, nor is the paper of Fischbach et al. (1986a) cited, the Eöt-Wash results did provide more stringent limits on the presence of such a force. It is difficult to make a direct comparison between the earlier and later results because the 1991 Eöt-Wash paper presented a limit on a force with a range of 30 m, whereas their 1994 paper gave limits for both 20 m and 50 m. The results are shown in Table 7.3. One can see that the uncertainty in the results has improved by a factor of approximately three and were consistent with no Fifth Force.

A group at the University of Zurich reported another test of the Fifth Force (Cornaz et al. 1994).²⁸ The experiment measured the difference in weight between two masses as a function of the height of the water in a pumped storage reservoir, Lake Gigerwald (Figure 7.17). “The basic idea of the Gigerwald experiment was to measure the weight difference of two test masses located above and below the variable water level with a single balance” (Cornaz et al. 1994, p. 1152). The experimental design avoided several of the problems of such experiments. “Since the weight difference is measured in a short time, balance drifts are negligible. Time-variable gravity effects originating from distances much larger than the separation of test masses completely vanish (e.g., tides). By comparing the weight differences

²⁶The title of the paper was, “New tests of the universality of free fall.”

²⁷The group also stated that, “We also test Weber’s claim that solar neutrinos scatter coherently from single crystals with cross sections $\sim 10^{23}$ times larger than the generally accepted value and rule out the existence of such cross sections” (Su et al. 1994, p. 3614). For a more detailed history of this episode, see Franklin (2010).

²⁸The major purpose of the experiment, as the title of the paper reveals, was to measure G , the gravitational constant.

Fig. 7.17 Schematic view of the Gigerwald experiment. From Cornaz et al. (1994).

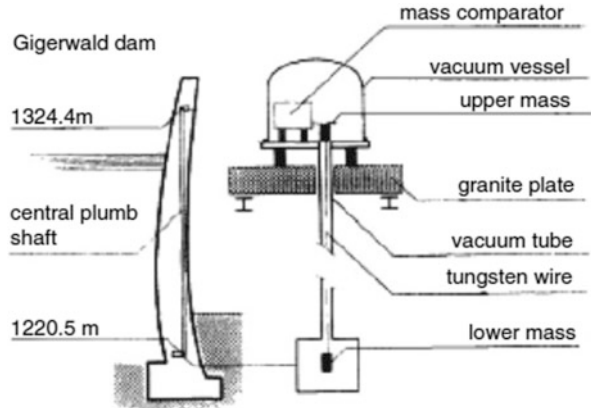
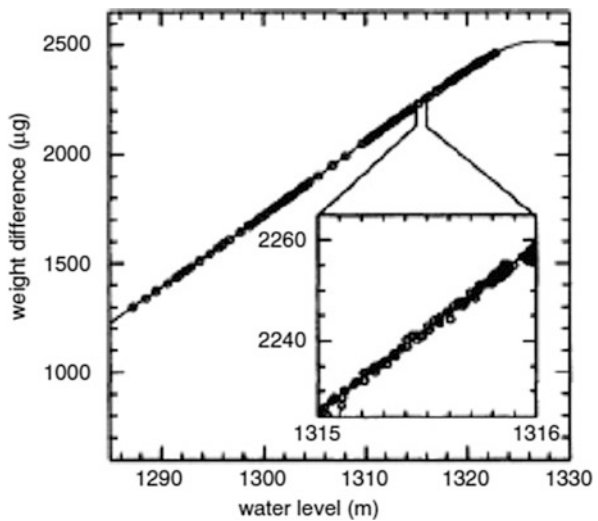


Fig. 7.18 The solid curve is the calculated weight difference of the two test masses as a function of the water level following pure Newtonian gravity (the origin is set at 1240 m for an empty lake). From Cornaz et al. (1994).

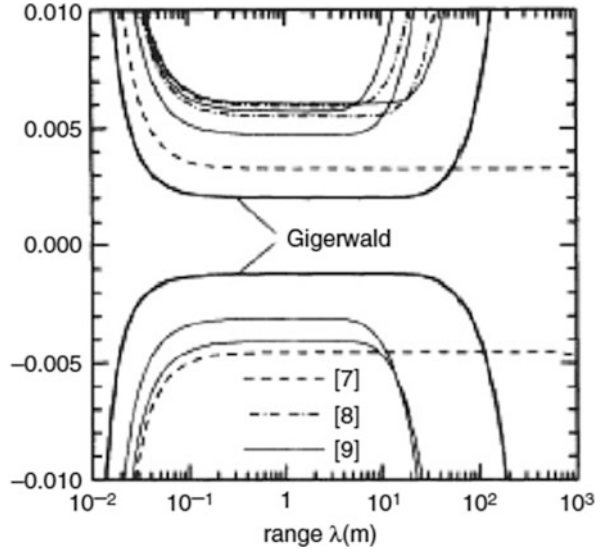


at several water levels even the static local gravity from the surroundings cancels. Finally, the recorded gravity signal is just due to the interaction between the locally moved mass (water and air) and the test masses” (Cornaz et al. 1994, pp. 1152–1153). The comparison between the theoretically calculated weight differences and the measured values is shown in Figure 7.18. The experimenters obtained more stringent limits on α , the strength of the proposed force, as a function of λ , the range, than had been obtained in previous experiments (Figure 7.19).²⁹

Experimental work on tests of the Fifth Force slowed, although there was still considerable theoretical work. In 1996 Carusotto et al. (1996) published their final results, which were the same as those discussed earlier, except for the inclusion of

²⁹This experiment was similar to those of Moore et al. (1988) and Bennett (1989).

Fig. 7.19 Excluded strengths α and ranges λ for a single Yukawa model at the 20 level arising from experiments measuring directly the gravitational constant at geophysical distances. From Cornaz et al. (1994).



a small systematic uncertainty. They concluded, “There is no evidence of any g -universality violation, at the level of μGal , at least with Galileo’s type experiment performed so far” Carusotto et al. (1996, p. 1274).

In 1997 Romaides et al. published their final results from the WABG tower experiment (Romaides et al. 1997). They had overcome the difficulties in making measurements at the largest height and stated, “...we succeeded in obtaining readings at 568 m above ground level. These readings, along with the previous results on the WABG and WTVD towers, allow for even tighter constraints on the non-Newtonian force parameters α and λ [the strength and range of the proposed Fifth Force]. Furthermore, we can now combine our tower data with data from lake experiments to give very tight constraints on the non-Newtonian coupling constant α over the entire geophysical window (10 m to 10 km)” (Romaides et al. 1997, p. 4352). Those constraints are shown in Figure 7.20. They concluded, “In summary, we conclude from existing tower experiments that at the present time there is no evidence for any significant deviation from the inverse-square law for $\lambda \approx 10^3$ m” (Romaides et al. 1997, p. 4356).

The Eöt-Wash group reported a new result using an interesting variant on their previous experimental apparatus (Gundlach et al. 1997). In their previous work, the group had used a torsion balance mounted on a rotating platform to measure the differential acceleration of various substances toward a local hillside and to other sources such as the Sun, the Earth, and the galaxy. In their latest experiment, the experimenters used a rotating three-ton ^{238}U attractor to measure the differential acceleration of lead and copper masses placed on a torsion balance. The Röt-Wash³⁰ apparatus is shown in Figure 7.21. The surroundings of the torsion balance

³⁰The Eöt-Wash group continued its whimsy with the naming of their new apparatus.

Fig. 7.20 Constraints on α in the range 10 m to 10 km. From Romaides et al. (1997).

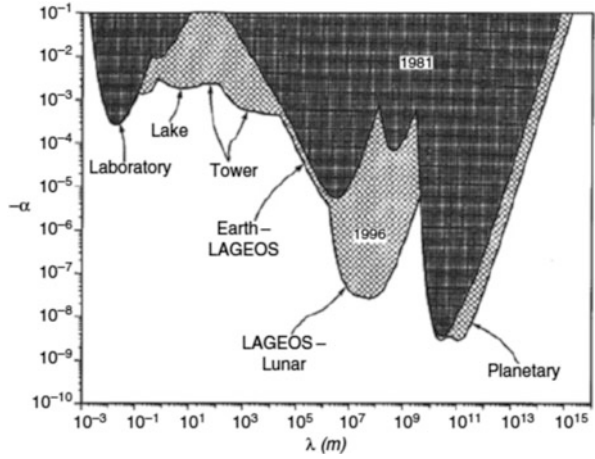
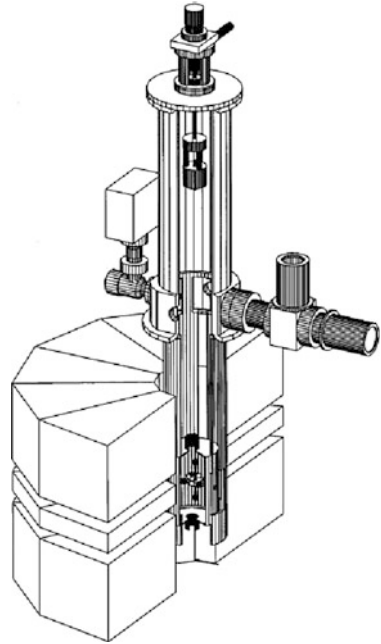


Fig. 7.21 Schematic view of the Röt-Wash instrument. The ^{238}U was counterbalanced by 820 kg of lead, so the floor would not tilt as the attractor revolved. From Gundlach et al. (1997).



were temperature controlled to guard against possible temperature effects. The ^{238}U was counterbalanced by 820 kg of lead, so the floor would not tilt as the attractor revolved. Tilt was a significant source of possible background effects in the Eöt-Wash experiments. The reason for the modification of the apparatus was that their previous experiment (Su et al. 1994) had been unable to test for forces with a range from 10 km to 1000 km. The new apparatus, using a local source, allowed such a test. The experimenters concluded, “ We found that $a_{\text{Cu}} - a_{\text{Pb}} = (-0.7 \pm 5.7) \times 10^{-13} \text{ cm/s}^2$, compared to the $9.8 \times 10^{-5} \text{ cm/s}^2$ gravitational acceleration toward the attractor. Our results set new constraints on equivalence-principle violating interactions with Yukawa ranges down to 1 cm and rule out an earlier suggestion of a Yukawa interaction coupled predominantly to $N - Z$ ” (Gundlach et al. 1997, p. 2523).

In 1997 George Gillies published a review of measurements of the gravitational constant and other related measurements. He remarked that, “The contemporaneous suggestion by Fischbach et al. (1986a) that there may be previously undiscovered, weak, long-range forces in nature provided further impetus for investigating the composition- and distance-dependence of gravity, since the presence of any such effect might reveal the existence of a new force. During this time, a theoretical framework for admitting non-Newtonian effects into discussions of the experimental results was emerging. It led to the practice of using the laboratory data to set limits on the size of the strength-range parameters in a Yukawa term added onto the Newtonian potential, and this has become a standard method for intercomparing the results of this class of experiments. Even though convincing evidence in favour of such new weak forces was never found, the many resulting experiments, when viewed as tests of the universality of free-fall, did much to improve the experimental underpinnings of the weak equivalence principle (WEP) of general relativity. In fact, searches for departures from the inverse square behaviour of Newtonian gravity have now come to be interpreted as attempts to uncover violations of the WEP” (Gillies 1997, p. 200).

After a decade of negative experimental results of the Fifth Force, 1997 produced a positive result. Achilli et al. (1997), using a superconducting gravimeter, measured changes in the gravitational force caused by the changing water level in a pumped storage reservoir, Lake Brasimone in Italy, and found evidence for a violation in the distance dependence of Newton’s law (Figure 7.22). The superconducting gravimeter could measure variations in gravity of the order of 1 nGal (1 Gal = 1 cm/s^2). A problem for the experimenters was the fact that tidal effects were of the order of 100–250 μGal . That effect could not be calculated precisely, so the group measured the lake tides for a period of 5 months at a location 400 m from the lake. The experimenters also obtained a detailed survey of the lake shore, an important factor in obtaining a result.

The gravimeter measured the gravitation effect by measuring the feedback force needed to maintain a levitated superconducting niobium sphere in a fixed position (Figure 7.23). They calibrated their apparatus by moving a known annular mass vertically with the gravimeter at its center. They also compared their gravimeter to an absolute gravimeter from another laboratory. The experimenters also investigated

Fig. 7.22 Sketch of the Lake Brasimone experiment. From Achilli et al. (1997).

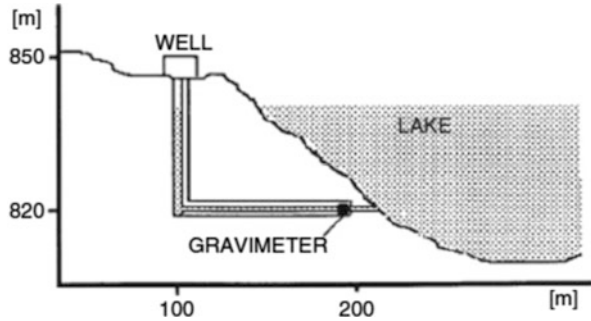
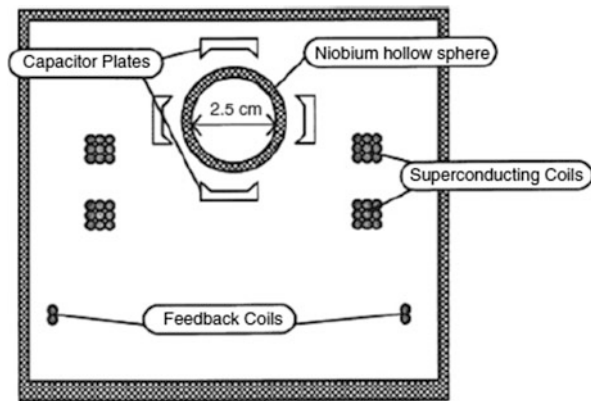


Fig. 7.23 Schematic cross-sectional view of the gravity sensor. The entire apparatus is contained in a liquid helium bath. From Achilli et al. (1997).

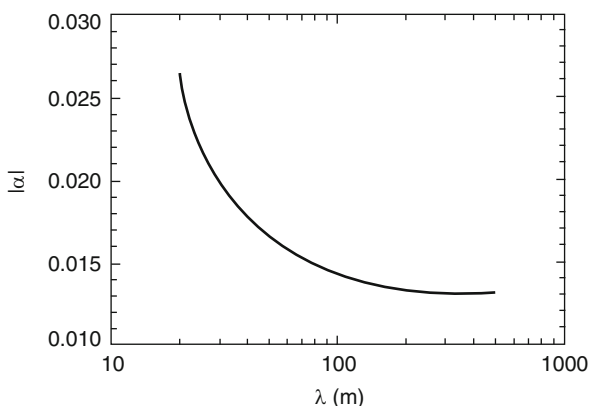


and measured geological, temperature, water table, and density background effects. Their final result $R = \text{observed/theoretical effect}$ was 1.0127 ± 0.0013 (Figure 7.24). “The ratio between the measured and expected gravitational effects differs from 1 by more than 9 standard deviations” (Achilli et al. 1997, p. 775). The experimenters noted, however, that, “. . . the only parameter not verified at the 0.1% level was the gravimeter calibration factor. In any case, the adopted value is in agreement with the result of the comparison with an absolute gravimeter” (Achilli et al. 1997, p. 802). Their results for $|\alpha|$ as a function of λ are shown in Figure 7.24. The group stated that their result differed from that found by Cornaz et al. (1994) in a similar experiment (see earlier discussion).

7.3.2 The Twenty-First Century

The Eöt-Wash group continued taking data with their rotating ^{238}U attractor. They remarked that, “Our new results set new constraints on equivalence principle violating interactions with Yukawa ranges down to 1 cm, and improved by substantial factors existing limit for ranges between 10 km and 1000 km” (Smith et al. 2000,

Fig. 7.24 $|\alpha|$ versus λ in the range 20 m to 500 m. From Achilli et al. (1997).



p. 022001-1). Their results are shown in Figure 7.25. Their new value for the difference in acceleration for copper and lead masses was $a_{\text{Cu}} - a_{\text{Pb}} = (-1.0 \pm 2.8) \times 10^{-13} \text{ cm/s}^2$, with the uncertainty reduced by a factor of two compared to their 1997 result.

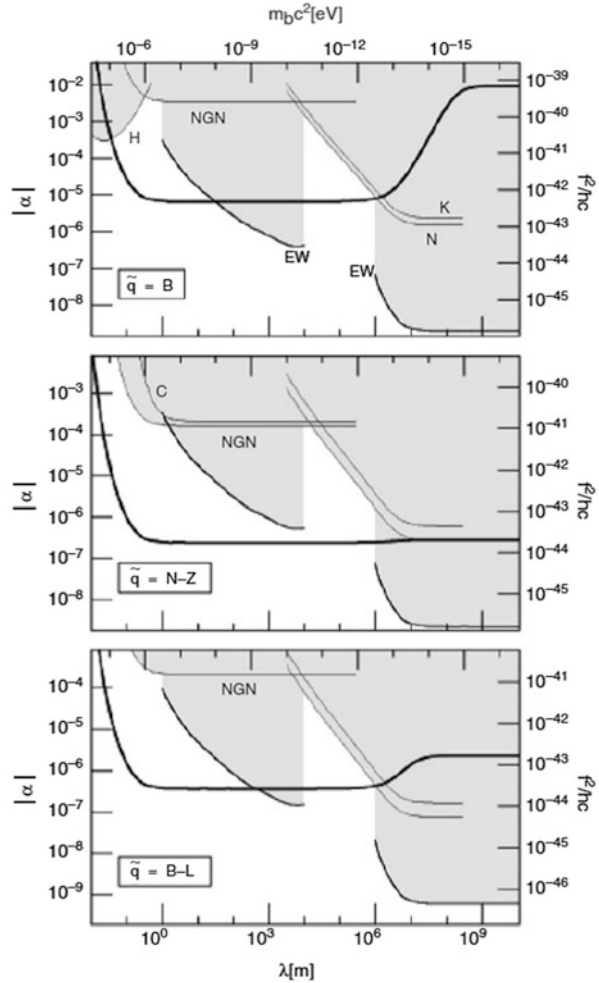
Perhaps the most interesting result reported in 2000 was the withdrawal of the positive Fifth Force result of Achilli et al. (1997). As Focardi, a member of the group remarked, “The above result [the positive result] convinced us of the importance of making any possible effort to check the conclusions reached in the previous experiment” (Focardi 2002, p. 419).³¹ This withdrawal was based on a reanalysis of the same data used in the 1997 paper. (A more detailed discussion of the reanalysis appeared in Baldi et al. 2001.) The experimenters performed a new and better calibration of their superconducting gravimeter and included a more consistent model of tidal gravity variations. Their initial paper had stated that “the only parameter not verified at the 0.1% level was the gravimeter calibration factor” (Achilli et al. 1997, p. 802). Their new result for $R = \text{experimental value/theoretical calculation} = 1.0023 \pm 0.0017$. This should be compared with their earlier result of $R = 1.023 \pm 0.0017$. They concluded that, “The result of this analysis shows an agreement between data and Newtonian theory to within 0.1 % level” (Baldi et al. 2001, p. 082001-2). At the turn of the twenty-first century, there was still no evidence supporting the Fifth Force.

In 2001 Bennett reported a second result from his experiment conducted at the Little Goose Lock on the Snake River. This was a torsion pendulum experiment which used the changing amount of water in the lock as an attractor. His initial data was taken in 1988 and published in 1989 (Bennett 1989). His 2001 paper included additional data taken in 1990.³²

³¹Focardi’s paper was presented at a conference in 2000, but the conference proceedings were not published until 2002.

³²For various personal reasons, Bennett did not publish these results until 2001.

Fig. 7.25 95% confidence limits on $|\alpha|$ vs λ for hypothetical interactions coupling to vector charges $q = B$, $q = N - Z$, or $q = B - L$, where B is baryon number, N is the number of nucleons in the nucleus, Z is the number of protons, and L is the number of leptons. The heavy curves are from this work. From Smith et al. (2000).



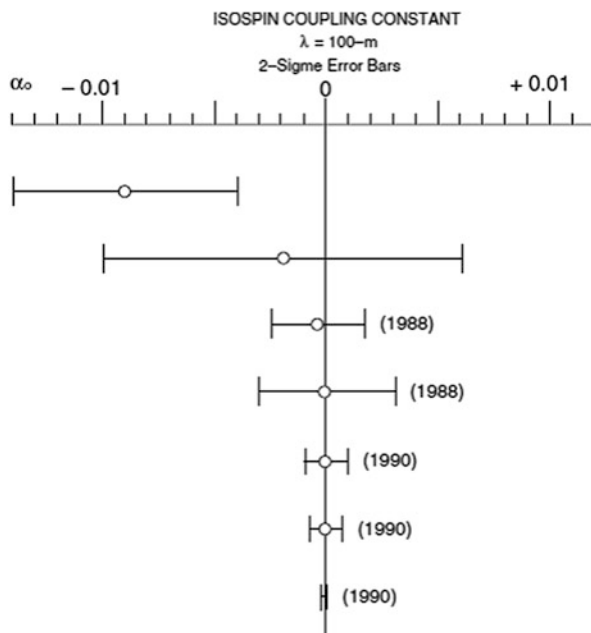
Bennett had made improvements in his apparatus including replacing the copper-lead disk in his torsion pendulum with a copper-lead annular ring. “A $2 - \sigma$ limit was set on the “isospin coupling constant” of $\alpha_0 = \pm 0.001$ at $\lambda = 100$ m” (Bennett 2001, p. 123). He also presented a summary of the $1 - \sigma$ limits on the differential acceleration for various pairs of substances (Table 7.4) along with a comparison of the coupling constants, α_0 , obtained by various experiments (Figure 7.26). The Fifth Force was still absent.

Despite the negative evidence, new experimental tests of the Fifth Force and of the weak equivalence principle were still being planned. Dittus and Mehls (2001), for example, were building a free-fall experiment in which two test masses of different substances would be dropped from a height of 110 m at the Bremen Tower. Any difference in fall would be detected by a SQUID (superconducting quantum

Table 7.4 Comparison of 1- σ limits on differential acceleration From Bennett (2001)

Reference	$\Delta a \times 10^{10} \text{cm/sec}^2$	Test masses	Source
(Thieberger 1987)	850 ± 260	Cu-H ₂ O	Cliff
(Fitch et al. 1988)	30 ± 49	Cu-CH ₂	Sloping terrain
(Bennett 1989)	25 ± 52	Cu-Pb	H ₂ O
(Bennett 2001)	2 ± 22	Cu-Pb	H ₂ O
(Adelberger et al. 1990)	-0.15 ± 2.6	Be-Al	Pb

Fig. 7.26 Comparison of different determinations of the intrinsic coupling coefficient α_0 for isospin coupling. From Bennett (2001); note that “BENNETT (1990)” is Bennett (2001).



interference device). They were aiming at an accuracy of better than 10^{-12} in the Eötvös ratio $\eta = 2((m_g/m_i)_1 - (m_g/m_i)_2)/((m_g/m_i)_1 + (m_g/m_i)_2)$, where m_i and m_g are the inertial and gravitational masses and the indices 1 and 2 are for the test masses of different substances. They remarked that the then current best value for η was less than 10^{-12} obtained by the Eöt-Wash group (Su et al. 1994).

Reasenber and Phillips were developing a different type of apparatus. “We are developing a Galilean test of the equivalence principle in which two pairs of test mass assemblies (TMA) are in free fall in a comoving vacuum chamber for about 0.9 s. The TMA are tossed upward, and the process repeats at 1.2 s intervals.³³ Each TMA carries a solid quartz retroreflector and a payload mass of about one-third of the total TMA mass. The relative vertical motion of the TMA of each

³³The title of their paper is “Testing the equivalence principle on a trampoline.”

pair is monitored by a laser gauge working in an optical cavity formed by the retroreflectors. Single-toss precision of the relative acceleration of a single pair of TMA is 3.5×10^{-12} g. The project goal of $\Delta g/g = 10^{-13}$ can be reached in a single night's run" (Reasenberg and Phillips 2001, p. 2435).

In 2002 as part of a proposed satellite experiment to test the weak equivalence principle, Moffat and Gillies summarized the current state of such tests. "In a long series of elegant experiments with rotating torsion balances, the Eöt-Wash Group has searched for composition dependence in the gravitational force via tests of the universality of free fall. In terms of the standard Eötvös parameter η , they have reached sensitivities of $\eta \sim 1.1 \times 10^{-12}$ in comparisons of the accelerations of Be and Al/Cu test masses and, more recently, have resolved differential accelerations of approximately 1.0×10^{-14} cm s⁻² in experiments with other masses. Drop-tower experiments now underway in Germany have as their goal testing WEP at sensitivities of $\eta \sim 1 \times 10^{-13}$, and Unnikrishnan describes a methodology under study at the Tata Institute of Fundamental Research in India wherein torsion balance experiments aiming at sensitivities of $\eta \sim 1 \times 10^{-14}$ are being developed" (Moffat and Gillies 2002, p. 92.3). None of these experiments provided evidence for the Fifth Force. The authors noted that proposed space-based experiments expected greater sensitivity. It was not clear, however, whether such experiments would cast any light on the Fifth Force, as initially proposed.

There were no other significant experimental tests of the Fifth Force in the early part of the twenty-first century. There were, however, experiments to measure G , the universal gravitational constant, a parameter whose value was then, and is now, uncertain. There were also experiments testing the law of gravity at very short distances, as well as continued discussions of space experiments. In 2005, Jens Gundlach, a member of the Eöt-Wash collaboration, published a review of the evidence to that date. His conclusion was, "At the moment, no deviations from ordinary gravity have been found, ..." (Gundlach 2005, p. 21). Faller (2005) published an amusing review of measurements of g , the acceleration due to gravity at the surface of the Earth. Faller and his collaborators had previously tested the Fifth Force hypothesis in both Galileo-type falling body experiments and by measuring g as a function of height in a tower. He noted that, "In the end (numerous experiments by many workers later), Newtonian gravity was vindicated" (Faller 2005, p. 571). He also related an amusing anecdote concerning the use of the tower in Erie, Colorado in the tests of the inverse-square law of gravity. "NOAA asked a modest \$1000 in rent for our use of the tower. Their other requirement was that we sign a paper to the effect that if we fell off in the course of making measurements, NOAA would not be held responsible for any personnel free falling due to gravity" (Faller 2005, p. 571).

The Eöt-Wash collaboration continued their extensive study of the equivalence principle with a new and improved torsion balance (Schlamminger et al. 2008). Their results for the difference in acceleration for beryllium and titanium test masses, in the northern and western directions, are shown in Figure 7.27. A violation of the equivalence principle would appear as a difference in the means of the runs taken with the masses in different orientations. The small offset was due to a systematic error, which did not affect their conclusion. Their new upper limits for α , the strength parameter for the Fifth Force or any other deviation from the

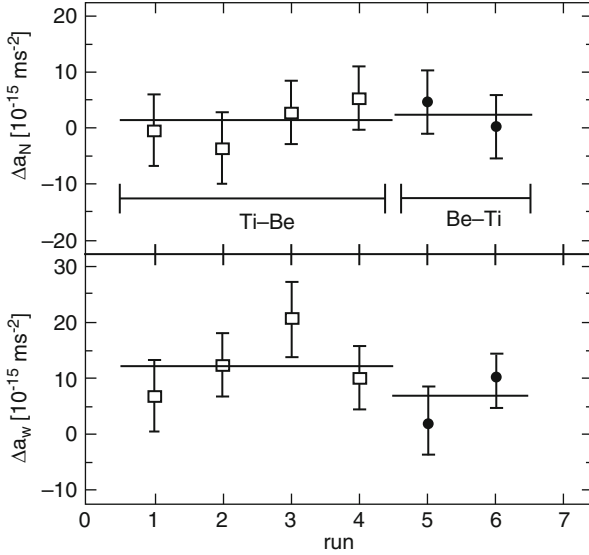
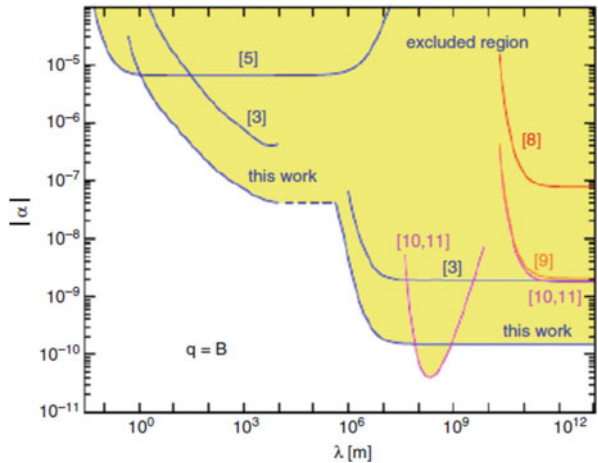


Fig. 7.27 Shown are measured differential accelerations toward north (top) and west. After the first four data runs, the Be and Ti test bodies were interchanged on the pendulum frame. A violation of the equivalence principle would appear as a difference in the means (lines) of the two data sets. The offset acceleration is due to systematic effects that follow the pendulum frame but not the composition dipole. From Schlamminger et al. (2008).

Fig. 7.28 New upper limits on Yukawa interactions coupled to baryon number with 95% confidence. From Schlamminger et al. (2008).



law of gravity, are shown in Figure 7.28. The region of interest for the Fifth Force is at approximately 100 m.³⁴ “We used a continuously rotating torsion balance instrument to measure the acceleration difference of beryllium and titanium test bodies towards sources at a variety of distances. Our result $\alpha_{N,Be-Ti} = (0.6 \pm 3.1) \times 10^{-15} \text{ m/s}^2$ improves limits on equivalence-principle violations with ranges from 1 m to ∞ by an order of magnitude. The Eötvös parameter is $\eta_{Earth,Be-Ti} = (0.3 \pm 1.8) \times 10^{-13}$ ” (Schlamminger et al. 2008, p. 041101-1). Recall that their previous best limit for η was 1.1×10^{-12} . The Fifth Force, if it existed, was becoming weaker.

In 2009 a review paper on torsion balance experiments by members of the Eöt-Wash group appeared (Adelberger et al. 2009; Gundlach et al. 2009). Adelberger and collaborators discussed details and experimental issues involved in torsion balance experiments as well as past experiments and proposed future experiments. The “Fifth Force” era received only a very brief summary. “After the completion of the classic experiments,³⁵ little further activity took place until 1986 when Fischbach et al. (1986a) reanalysed the Eötvös data. They used this, along with previous claims of anomalous data on g in mines, to claim evidence for a new force. This “Fifth Force” was an EP-violating acceleration coupled to B with a range of a few hundred meters that would have rendered it invisible to the classic solar EP tests. This finding triggered many experiments looking for intermediate-range ($10 \text{ m} < \lambda < 10000 \text{ km}$) forces. The Eöt-Wash group at the University of Washington responded by developing a torsion balance mounted on a uniformly rotating platform . . . The first result from this instrument, which appeared in 1987, ruled out the original Fifth Force proposal.³⁶ However, the suggestion of a finite-ranged Yukawa interaction led physicists to broaden their view of EP tests to a search for Yukawa interactions at all accessible length scales” (Adelberger et al. 2009, pp. 108–109).

After 2010 there was very little experimental activity that explicitly dealt with the Fifth Force. This is not to say that there was no work on the related topic of the universality of free fall and tests of the weak equivalence principle. Various experiments conducted in space tested that principle at distances larger than the range of the Fifth Force, and there were laboratory experiments that investigated the law of gravity at much smaller distances. An entire issue of *Classical and Quantum Gravity* (Volume 29, Issue 18, 2012) was devoted solely to tests of the weak equivalence principle. The Eöt-Wash group paper in that volume reported a new result (Wagner et al. 2012). In addition to their previous result of $\alpha_{N,Be-Ti} =$

³⁴This was the approximate range suggested in the initial paper, based on the (later withdrawn) results of Stacey and his collaborators. The data of the Eötvös and his collaborators is consistent with ranges up to 1 AU.

³⁵These were the experiments which test the weak equivalence principle in the fall of bodies toward the Sun: Braginskii and Panov (1972) and Roll et al. (1964).

³⁶As we saw in Section 7.2 and in the history presented above, this is not accurate.

$(0.6 \pm 3.1) \times 10^{-15} \text{ m/s}^2$, they presented a new result for an aluminum-beryllium pair, $\alpha_{\text{N,Be-AL}} = (-1.2 \pm 2.2) \times 10^{-15} \text{ m/s}^2$.

Will (2014) summarized the situation with respect to the Fifth Force in an extensive review, “The Confrontation between General Relativity and Experiment.” He concluded that, “A consensus emerged that there was no credible evidence for a fifth force of nature, of a type and range proposed by Fischbach et al.” (Will 2014, p. 27). Will’s summary is, as we have seen, accurate.

7.3.3 Discussion

There is very strong and persuasive evidence that the Fifth Force, as initially proposed by Ephraim Fischbach and his collaborators, does not exist. Numerous experiments have not shown the presence of any force with strength approximately one percent that of Newtonian gravity and with a range of about 100 m. I believe, however, that the hypothesis has been quite fruitful. It encouraged renewed interest in tests of general relativity, particularly on the weak equivalence principle and on Newtonian gravity at both very large and very small distances and on its composition dependence. This work also led to improvements in both experimental apparatuses and experimental analyses. As Gillies remarked in 1997, “The contemporaneous suggestion by Fischbach et al. (1986a) that there may be previously undiscovered, weak, long-range forces in nature provided further impetus for investigating the composition- and distance-dependence of gravity, since the presence of any such effect might reveal the existence of a new force Even though convincing evidence in favour of such new weak forces was never found, the many resulting experiments, when viewed as tests of the universality of free-fall, did much to improve the experimental underpinnings of the weak equivalence principle (WEP) of general relativity. In fact, searches for departures from the inverse square behaviour of Newtonian gravity have now come to be interpreted as attempts to uncover violations of the WEP” (Gillies 1997, p. 200).

Some scholars have suggested that the Fifth Force hypothesis should never have been further investigated (Anderson 1992). These after-the-fact judgments are, I believe, incorrect. As mentioned above the hypothesis was quite fruitful. In addition, I believe that it is important to recognize that wrong science is not bad science. The fact that the Fifth Force hypothesis turned out to be incorrect is not a good reason for saying that it should not have been further investigated. There was, at the time, plausible evidence from the reanalysis of the Eötvös experiment, from the discrepancy between laboratory and mineshaft measurements of g and from the tantalizing energy dependence of the K^0 decay parameters that was consistent with the hypothesis. Although one might argue that it was an unlikely hypothesis, the history of science has shown that on occasion such hypotheses have turned out to be correct. Consider the case of parity nonconservation. Distinguished scientists such as Wolfgang Pauli and Richard Feynman were willing to bet that the suggestion by

Lee and Yang that parity was not conserved in the weak interactions was incorrect. Feynman bet Norman Ramsey 50 to 1 that parity would be conserved. When experiments showed that parity was not conserved, Feynman paid (for details see Franklin 1986, Chapter 1).

The episode of the Fifth Force is an illustration of good science. A speculative hypothesis, one with some evidential support, was proposed. Further experimentation demonstrated that the hypothesis was incorrect. It did, however, lead to further experimental and theoretical work and improvements in experiments.

References

- Achilli, V., Baldi, P., et al. (1997). A geophysical experiment on Newton's inverse-square law. *II Nuovo Cimento*, 112 B, 775–803.
- Adelberger, E. G. (1988). Constraints on composition-dependent interactions from the Eöt-wash experiment. In O. Fackler & J. Tran Thanh Van (Eds.), *5th Force Neutrino Physics: Eighth Noriond Workshop*. Gif sur Yvette: Editions Frontières.
- Adelberger, E. G. (1989). High-sensitivity hillside results from the Eöt-wash experiment. In O. Fackler & J. Tran Thanh Van (Eds.), *Tests of Fundamental Laws in Physics: Ninth Moriond Workshop* (pp. 485–499). Les Arcs, France: Editions Frontières.
- Adelberger, E. G., Stubbs, C. W., Rogers, W. F., et al. (1987). New constraints on composition-dependent interactions weaker than gravity. *Physical Review Letters*, 59, 849–852.
- Adelberger, E. G., Stubbs, C. W., et al. (1990). Testing the equivalence principle in the field of the earth: Particle physics at masses below 1μ eV. *Physical Review D*, 42, 3267–3292.
- Adelberger, E. G., et al. (1991). Searches for new macroscopic forces. *Annual Review of Nuclear and Particle Science*, 41, 269–320.
- Adelberger, E. G., Gundlach, J. H., et al. (2009). Torsion balance experiments: A low-energy frontier of particle physics. *Progress in Particle and Nuclear Physics*, 62, 102–134.
- Ander, M., Zumbege, M. A., Lautzenhiser, T., et al. (1989). Test of Newton's inverse-square law in the Greenland ice cap. *Physical Review Letters*, 62, 985–988.
- Anderson, P. (1992). The reverend Thomas Bayes, needles in haystacks, and the fifth force. *Physics Today*, 45, 9–11.
- Aronson, S. H., Bock, G. J., Cheng, H. Y., et al. (1983a). Energy dependence of the fundamental parameters of the $K^0 - \bar{K}^0$ system. I. Experimental analysis. *Physical Review D*, 28, 476–494.
- Aronson, S. H., Bock, G. J., Cheng, H. Y., et al. (1983b). Energy dependence of the fundamental parameters of the $K^0 - \bar{K}^0$ system. II. Theoretical formalism. *Physical Review D*, 28, 494–523.
- Baldi, P., Campari, E. G., et al. (2001). Testing Newton's inverse square law at intermediate scales. *Physical Review D*, 64, 082001-1–082001-7.
- Bartlett, D. F., & Tew, W. L. (1989a). The fifth force: Terrain and pseudoterrain. *Tests of Fundamental Laws in Physics: Ninth Moriond Workshop*. Les Arcs, France: Editions Frontières.
- Bartlett, D. F., & Tew, W. L. (1989b). Possible effect of the local terrain on the Australian fifth-force measurement. *Physical Review D*, 40, 673–675.
- Bartlett, D. F., & Tew, W. L. (1990). Terrain and geology near the WTVD tower in North Carolina: Implications for non-Newtonian gravity. *Journal of Geophysical Research*, 95, 363–369.
- Bell, J. S., & Perring, J. (1964). 2π decay of the K_2^0 meson. *Physical Review Letters*, 13, 348–349.
- Bennett, W. R. (1989). Modulated-source Eötvös experiment at little goose lock. *Physical Review Letters*, 62, 365–368.
- Bennett, W. R. (2001). Hunting the fifth force on the Snake River. In D. Budker, P. H. Bucksbaum & S. J. Freedman (Eds.), *Art and symmetry in experimental physics* (Vol. 596, pp. 123–155). Berkeley, CA: American Institute of Physics.

- Bernstein, J., Cabibbo, N., & Lee, T. D. (1964). CP invariance and the 2π decay of the K_2^0 . *Physics Letters*, 12, 146–148.
- Bizzeti, P. G. (1986). Significance of the Eötvös method for the investigation of intermediate range forces. *Il Nuovo Cimento*, 94B, 80–86.
- Bizzeti, P. G., Bizzeti-Sona, A. M., Fazzini, T., et al. (1988). New search for the ‘fifth force’ with the floating body method: Status of the Vallambrosa experiment. In O. Fackler & J. Tran Thanh Van (Eds.), *Fifth Force Neutrino Physics: Eighth NMoriond Workshop*. Gif Sur Yvette: Editions Frontières.
- Bizzeti, P. G., Bizzeti-Sona, A. M., Fazzini, T., et al. (1989a). Search for a composition dependent fifth force: Results of the Vallambrosa experiment. In J. Tran Thanh Van, & O. Fackler (Eds.), *Proceedings of the XXIVth Rencontre de Moriond* (pp. 511–524). Gif Sur Yvette: Editions Frontières.
- Bizzeti, P. G., Bizzeti-Sona, A. M., Fazzini, T., et al. (1989b). Search for a composition-dependent fifth force. *Physical Review Letters*, 62, 2901–2904.
- Bock, G. J., Aronson, S. H., Freudenreich, K., et al. (1979). Coherent KS regeneration by protons from 30 to 130 GeV/c. *Physical Review Letters*, 42, 350–353.
- Boynton, P. (1990). New limits on the detection of a composition-dependent macroscopic force. In O. Fackler & J. Tran Thanh Van (Eds.), *New and Exotic Phenomena ‘90: Tenth Moriond Workshop* (pp. 207–224). Gif sur Yvette: Editions Frontières.
- Boynton, P., & Peters, P. (1989). Torsion pendulums, fluid flows and the coriolis force. In O. Fackler and J. Tran Thanh Van (Eds.), *Tests of Fundamental Laws in Physics: Ninth Moriond Workshop* (pp. 501–510). Gif sur Yvette: Editions Frontières.
- Boynton, P., Crosby, D., Ekstrom, P., et al. (1987). Search for an intermediate-range composition-dependent force. *Physical Review Letters*, 59, 1385–89.
- Braginskii, V. B., & Panov, V. I. (1972). Verification of the equivalence of inertial and gravitational mass. *JETP Letters*, 34, 463–466.
- Brans, C., & Dicke, R. H. (1961). Mach’s principle and a relativistic theory of gravitation. *Physical Review*, 124, 925–935.
- Carusotto, S., Cavasinni, V., et al. (1992). Test of g universality with a Galileo type experiment. *Physical Review Letters*, 69, 1722–1725.
- Carusotto, S., Cavasinni, V., et al. (1993). Limits on the violation of G -universality with a Galileo-type experiment. *Physics Letters A*, 183, 355–358.
- Carusotto, S., Cavasinni, V., et al. (1996). g -Universality test with a Galileo’s type experiment. *Il Nuovo Cimento*, 111 B, 1259–1275.
- Christenson, J. H., Cronin, J. W., Fitch, V. L., et al. (1964). Evidence for the 2π decay of the K_2^0 meson. *Physical Review Letters*, 13, 138–140.
- Colella, R., Overhauser, A. W., & Werner, S. A. (1975). Observations of gravitationally induced quantum interference. *Physical Review Letters*, 34, 1472–1474.
- Cooper, L. (1935). *Aristotle, Galileo, and the Tower of Pisa*. Ithaca: Cornell University Press.
- Cornaz, A., Hubler, B., et al. (1994). Determination of the gravitational constant at an effective interaction distance of 112 m. *Physical Review Letters*, 72, 1152–1155.
- Cowsik, R., Krishnan, N., Tandor, S. N., et al. (1988). Limit on the strength of intermediate-range forces coupling to isospin. *Physical Review Letters*, 61, 2179–2181.
- Cruz, J. Y., Harrison, J. C., et al. (1991). A test of Newton’s inverse square law of gravitation using the 300 m tower at Erie, Colorado. *Journal of Geophysical Research*, 96, 20073–20092.
- De Rujula, A. (1986a). Are there more than four? *Nature*, 323, 760–761.
- De Rujula, A. (1986b). On weaker forces than gravity. *Physics Letters*, 180 B, 213–220.
- DeBouard, X., Dekkers, D., Jordan, B., et al. (1965). Two pion decay of the K_2^0 at 10 GeV/c. *Physics Letters*, 15, 58–61.
- Dittus, H., & Mehls, C. (2001). A new experimental baseline for testing the weak equivalence principle at the Bremen drop tower. *Classical and Quantum Gravity*, 18, 2417–2425.
- Eckhardt, D. H. (1986). Comment on ‘Reanalysis of the Eötvös experiment.’ *Physical Review Letters*, 57, 2868.

- Eckhardt, D. H. (1988). Tower gravity experiment: Evidence for non-Newtonian gravity. *Physical Review Letters*, 60, 2567–2570.
- Eckhardt, D. H., Jekeli, C., Lazarewicz, A. R., et al. (1988). Results of a tower gravity experiment. In O. Fackler & J. Tran Thanh Van (Eds.), *Fifth force neutrino physics: Eighth Moriond workshop* (pp. 577–583). Gif sur Yvette: Editions Frontières.
- Eckhardt, D.H., Jekeli, C., Lazarewicz, A. R., et al. (1989). Evidence for non-Newtonian gravity: Status of the AFGL experiment January 1989. *Tests of Fundamental Laws in Physics: Ninth Moriond Workshop*, Les Arcs, France: Editions Frontières.
- Eötvös, R., Pekar, D., & Fekete, E. (1922). Beiträge zum Gesetze der Proportionalität von Trägheit und Gravitation. *Annalen der Physik (Leipzig)*, 68, 11–66.
- Fairbank, W. M. (1988). Summary talk on fifth force papers. In O. Fackler & J. Tran Thanh Van (Eds.), *5th Force Neutrino Physics: Eighth Moriond Workshop* (pp. 629–644). Gif sur Yvette: Editions Frontières.
- Faller, J. E. (2005). The measurement of little g : A fertile ground for precision measurement science. *Journal of Research of the National Institute of Standards and Technology*, 110, 559–581.
- Fischbach, E. (1980). In P. Bergmann & V. De Sabbata (Eds.), *Tests of general relativity at the quantum level. Cosmology and gravitation* (pp. 359–373). New York: Plenum.
- Fischbach, E., & Freeman, B. (1979). Testing general relativity at the quantum level. *General Relativity and Gravitation*, 11, 377–381.
- Fischbach, E., & Talmadge, C. (1992). Six years of the fifth force. *Nature*, 356, 207–215.
- Fischbach, E., Aronson, S., Talmadge, C., et al. (1986a). Reanalysis of the Eötvös experiment. *Physical Review Letters*, 56, 3–6.
- Fischbach, E., Aronson, S., Talmadge, C., et al. (1986b). Response to Thodberg. *Physical Review Letters*, 56, 2424.
- Fitch, V. L., Isaila, M. V., & Palmer, M. A. (1988). Limits on the existence of a material-dependent intermediate-range force. *Physical Review Letters*, 60, 1801–1804.
- Focardi, S. (2002). The Newton's gravitational law. In R. Cianci, R. Collina, M. Francaviglia, & P. Fre (Eds.), *Recent developments in general relativity, Genoa 2000* (pp. 417–421). Genoa: Springer.
- Franklin, A. (1986). *The neglect of experiment*. Cambridge: Cambridge University Press.
- Franklin, A. (1993). *The rise and fall of the fifth force: Discovery, pursuit, and justification in modern physics*. New York: American Institute of Physics.
- Franklin, A. (2010). Gravity waves and neutrinos: The later work of Joseph Weber. *Perspectives on Science*, 18, 119–151.
- Franklin, A., & Fischbach, E. (2016). *The rise and fall of the fifth force*. Heidelberg: Springer.
- Fujii, Y. (1971). Dilatational possible non-Newtonian gravity. *Nature*, 234, 5–7.
- Fujii, Y. (1972). Scale invariance and gravity of hadrons. *Annals of Physics (N.Y.)*, 69, 494–521.
- Fujii, Y. (1974). Scalar-tensor theory of gravitation and spontaneous breakdown of scale invariance. *Physical Review D*, 9, 874–876.
- Galbraith, W., Manning, G., Taylor, A. E., et al. (1965). Two-pion decay of the K_2^0 meson. *Physical Review Letters*, 14, 383–386.
- Gibbons, G. W., & Whiting, B. F. (1981). Newtonian gravity measurements impose constraints on unification theories. *Nature*, 291, 636–638.
- Gillies, G. T. (1997). The Newtonian gravitational constant: Recent measurements and related studies. *Reports on Progress in Physics*, 60, 151–225.
- Gundlach, J. H. (2005). Laboratory test of gravity. *New Journal of Physics*, 7, 205-1–205-22.
- Gundlach, J. H., Schlamminger, S., et al. (2009). Laboratory tests of the equivalence principle at the University of Washington. *Space Science Review*, 148, 201–216.
- Gundlach, J. H., Smith, G. L., et al. (1997). Short-range test of the equivalence principle. *Physical Review Letters*, 78, 2523–2526.
- Heckel, B. R., Adelberger, E. G., Stubbs, C. W., et al. (1989). Experimental bounds on interactions mediated by ultralow-mass bosons. *Physical Review Letters*, 63, 2705–2708.

- Holding, S. C., Stacey, F. D., & Tuck, G. J. (1986). Gravity in mines—An investigation of Newton's law. *Physical Review D*, *33*, 3487–3494.
- Jekeli, C., Eckhardt, D. H., & Romaides, A. J. (1990). Tower gravity experiment: No evidence for non-Newtonian gravity. *Physical Review Letters*, *64*, 1204–1206.
- Kasameyer, P., Thomas, J., Fackler, O., et al. (1989). A test of Newton's law of gravity using the BREN tower, Nevada. *Tests of Fundamental Laws in Physics: Ninth Moriond Workshop*, Les Arcs, France: Editions Frontières.
- Kuroda, K., & Mio, N. (1989a). Galilean test for composition-dependent force. In D. G. Blair & M. J. Buckingham (Eds.), *Proceedings of the fifth Marcel Grossman conference on general relativity* (pp. 1569–1572). Singapore: World Scientific.
- Kuroda, K., & Mio, N. (1989b). Test of a composition-dependent force by a free-fall interferometer. *Physical Review Letters*, *62*, 1941–1944.
- Kuroda, K., & Mio, N. (1990). Limits on a possible composition-dependent force by a Galilean experiment. *Physical Review D*, *42*, 3903–3907.
- Liu, Y.-C., Yang, X.-S., et al. (1992). Testing non-Newtonian gravitation on a 320 m tower. *Physics Letters A*, *169*, 131–133.
- Long, D. R. (1974). Why do we believe Newtonian gravitation at laboratory dimensions? *Physical Review D*, *9*, 50–52.
- Lusignoli, M., & Pugliese, A. (1986). Hyperphotons and K-meson decays. *Physics Letters*, *171B*, 468–470.
- Maddox, J. (1986). Newtonian gravity corrected. *Nature*, *319*, 173.
- Mikkelsen, D. R., & Newman, M. J. (1977). Constraints on the gravitational constant at large distances. *Physical Review D*, *16*, 919–926.
- Milgrom, M. (1986). On the use of Eötvös-type experiments to detect medium-range forces. *Nuclear Physics*, *227B*, 509–512.
- Moffat, J. W., & Gillies, G. T. (2002). Satellite Eötvös test of the weak equivalence principle for zero-point vacuum energy. *New Journal of Physics*, *4*, 92.1–92.6.
- Moore, G. I. et al. (1988). Determination of the gravitational constant at an effective mass separation of 22 m. *Physical Review D*, *38*, 1023–1029.
- Nelson, P. G., Graham, D. M., & Newman, R. D. (1990). Search for an intermediate-range composition-dependent force coupling to N-Z. *Physical Review D*, *42*, 963–976.
- Neufeld, D. A. (1986). Upper limit on any intermediate-range force associated with Baryon number. *Physical Review Letters*, *56*, 2344–2346.
- Newman, R., Graham, D., & Nelson, P. (1989). A 'fifth force' search for differential acceleration of lead and copper toward lead. In O. Fackler & J. Tran Thanh Van (Eds.), *Tests of fundamental laws in physics: Ninth Moriond workshop* (pp. 459–472). Gif sur Yvette: Editions Frontières.
- Niebauer, T. M., McHugh, M. P., & Faller, J. E. (1987). Galilean test for the fifth force. *Physical Review Letters*, *59*, 609–612.
- Parker, R. L., & Zumbege, M. A. (1989). An analysis of geophysical experiments to test Newton's law of gravity. *Nature*, *342*, 29–32.
- Raab, F. J. (1987). Search for an intermediate-range interaction: Results of the Eöt-wash. I experiment. In O. Fackler & J. Tran Thanh Van (Eds.), *New and exotic phenomena: Seventh Moriond workshop*. Les Arcs, France: Editions Frontières: 567–577.
- Reasenber, R. D., & Phillips, J. D. (2001). Testing the equivalence principle on a trampoline. *Classical and Quantum Gravity*, *18*, 2435–2445.
- Roll, P. G., Krotkov, R., et al. (1964). The equivalence of inertial and passive gravitational mass. *Annals of Physics*, *26*, 442–517.
- Romaides, A. J., Sands, R. W., et al. (1994). Second tower experiment: Further evidence for Newtonian gravity. *Physical Review D*, *50*, 3608–3613.
- Romaides, A. J., Sands, R. W., et al. (1997). Final results from the WABG tower gravity experiment. *Physical Review D*, *55*, 4532–4536.
- Schlaminger, S., Choi, K. Y., et al. (2008). Test of the equivalence principle using a rotating torsion balance. *Physical Review Letters*, *100*, 041101-1–041101-4.

- Schwarzschild, B. (1986). Reanalysis of old Eötvös data suggests 5th force ... to some. *Physics Today*, 39(10): 17–20.
- Smith, G. L., Hoyle, C. D., et al. (2000). Short-range tests of the equivalence principle. *Physical Review D*, 61, 022001-1–022001-20.
- Speake, C. C., Niebauer, T. M., McHugh, M. P., et al. (1990). Test of the inverse-square law of gravitation using the 300-m tower at Erie Colorado. *Physical Review Letters*, 65, 1967–1971.
- Stacey, F. D., & Tuck, G. J. (1981). Geophysical evidence for non-Newtonian gravity. *Nature*, 292, 230–232.
- Stacey, F. D., Tuck, G. J., Holding, S. C., et al. (1981). Constraint on the planetary scale value of the Newtonian gravitational constant from the gravity profile with a mine. *Physical Review D*, 23, 1683–1692.
- Stubbs, C. W. (1990). Seeking new interactions: An assessment and overview. *New and exotic phenomena '90: Tenth Moriond workshop*. Les Arcs, France: Editions Frontières.
- Stubbs, C. W., Adelberger, E. G., Heckel, B. R., et al. (1989). Limits on composition-dependent interactions using a laboratory source: Is there a 'fifth force?' *Physical Review Letters*, 62, 609–612.
- Su, Y., Heckel, B. R., et al. (1994). New tests of the universality of free fall. *Physical Review D*, 50, 3614–3636.
- Thieberger, P. (1986). Hypercharge fields and Eötvös-type experiments. *Physical Review Letters*, 56, 2347–2349.
- Thieberger, P. (1987). Search for a substance-dependent force with a new differential accelerometer. *Physical Review Letters*, 58, 1066–1069.
- Thieberger, P. (1989). Thieberger replies. *Physical Review Letters*, 62, 810.
- Thodberg, H. H. (1986). Comment on the sign in the reanalysis of the Eötvös experiment. *Physical Review Letters*, 56, 2423.
- Thomas, J., Vogel, P., & Kasameyer, P. (1988). Gravity anomalies at the Nevada test site. In *5th force, neutrino physics: Eight Moriond workshop*. Les Arcs, France: Editions Frontières.
- Unnikrishnan, C. S. (1993). Search for a 5th force. *Pramana Journal of Physics*, 41(Supplement S), 395–411.
- Wagner, T. A., Schlamminger, S., et al. (2012). Torsion-balance tests of the weak equivalence principle. *Classical and Quantum Gravity*, 29(18), 184002-1–184002-15.
- Weinberg, S. (1964). Do hyperphotons exist? *Physical Review Letters*, 13, 495–497.
- Will, C. (1981). *Theory and experiment in gravitational physics*. Cambridge: Cambridge University Press.
- Will, C. (1984). *Was Einstein right?* New York: Basic Books.
- Will, C. W. (2014). The confrontation between general relativity and experiment. *Living Reviews in Relativity*, 17, 1–117.
- Yang, X., Liu, W., et al. (1991). Testing the intermediate-range force at separations around 50 meters. *Chinese Physics Letters*, 8, 329–332.

Part III
Geometry and Cosmology, Past and
Present

Chapter 8

Cyclic Models of the Relativistic Universe: The Early History



Helge Kragh

8.1 Introduction

Although the general idea of a cyclic or oscillating universe goes back to times immemorial, it was only with the advent of relativistic cosmology that it could be formulated in a mathematically precise way and confronted with observations. Ever since Friedmann (1922) introduced the possibility of a closed cyclic universe, it has continued to attract interest among a minority of astronomers and physicists. At the same time, it has been controversial and widely seen as speculative, in part because of its historical association with an antireligious world view. According to Steven Weinberg, “the oscillating model . . . nicely avoids the problem of Genesis” and may be considered philosophically appealing for that reason (Weinberg 1977, 154). In spite of many problems and a generally bad reputation, cyclic models never vanished from the scene of cosmology. Indeed, they have recently experienced a remarkable revival, especially in forms inspired by string theory (Steinhardt and Turok 2002, 2007).

The present essay covers the history of the cyclic universe, understood as a class of solutions to the cosmological field equations, in the period from 1922 to about 1960. No attempt is made to extend the investigation to the later development (which is covered in part by Kragh 2009). As the history of this kind of cosmological view goes up to the present, so it goes very far back in time, if more as a philosophical than a scientific idea. This earlier history is not part of my essay either, but contrary to the modern history, it is thoroughly described in the literature of history of science and ideas, e.g. Jaki (1974) and Kragh (2008).

H. Kragh (✉)

Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, 2100 Copenhagen, Denmark
e-mail: helge.kragh@nbi.ku.dk

8.2 The Friedmann-Einstein Universe

In his seminal work of 1922, the very beginning of evolutionary relativistic cosmology, Alexander Friedmann analysed from a mathematical perspective the various solutions to Einstein's field equations, including the possibility of a zero or negative cosmological constant (Friedmann 1922). His analysis rested on very general assumptions, namely, that the cosmic matter is at rest and exerts no pressure and also that the constant curvature space is orthogonal to time. What would later be known as the Friedmann equations appeared in the form

$$\left(\frac{R'}{R}\right)^2 + \frac{2RR'}{R^2} + \frac{c^2}{R^2} - \Lambda = 0 \quad (8.1)$$

and

$$3\left(\frac{R'}{R}\right)^2 + \frac{3c^2}{R^2} - \Lambda = \kappa c^2 \rho \quad (8.2)$$

where $\kappa = 8\pi G/c^2$ and ρ denotes the mean density of matter. Introducing a constant A given by the total mass M of the universe according to

$$A = \frac{\kappa M}{6\pi^2} = \frac{4GM}{3\pi c^2} \quad (8.3)$$

he showed that, in the case $\Lambda < 4c^2/9A^2$, the radius of curvature would become a periodic function of t with a “world period” given by

$$t_\pi = \frac{2}{c} \int_0^{x_0} \sqrt{\frac{x}{A - x + \frac{\Lambda x^3}{3c^2}}} dx \quad (8.4)$$

Friedmann commented: “The radius of curvature varies between 0 and x_0 . We shall call this universe the *periodic world*. The period of the periodic world increases if we increase Λ and tends towards infinity if Λ tends towards the value $\Lambda_1 = 4c^2/9A^2$ ” (Friedmann 1922, 385).¹ He further noted that for small values of Λ , the world period is given by

$$t_\pi \cong \frac{\pi A}{c} = \frac{\kappa M}{6\pi} \quad (8.5)$$

¹For the cosmological constant, I have substituted the symbol Λ for Friedmann's λ (which is also the symbol Einstein used in his article of 1917). An English translation of Friedmann's paper appears in several versions, e.g. Lang and Gingerich (1979, 838–843).

As an illustration, he calculated that for $\Lambda = 0$ and a world mass of $M = 5 \times 10^{21}$ solar masses t_π became of the order of 10 billion years. It is unknown why he chose this particular value of M (but see Tropp et al. (1993, 159) and also the comment by Georg Singer in Friedmann 2000, 137–138). In his paper of 1922, Friedmann did not describe the universe as oscillating in the sense that cycle followed after cycle but only referred to a single cycle from bang to crunch. Nor did he refer to thermodynamical or other physical properties such as the content of matter and radiation. His work was basically a mathematical investigation.

It is worth noting that Friedmann, after having introduced the idea of a world cycle, pointed out that the notion could be understood in two different ways. Two events could be counted as coincident if they have the same spatial coordinates at times t' and $t' \pm nt_\pi$ ($n = 1, 2, \dots$), which corresponds to the ordinary picture of a pulsating universe limited in time between $t = 0$ and $t = t_\pi$. Alternatively, “if the time varies between $-\infty$ and $+\infty$ (e.g. if we consider two events as coincident only when not only their spatial but also their world coordinates coincide), we come to a real periodicity of the space curvature” (Friedmann 1922, 385). Without elaborating, he adopted the first viewpoint. The second one involves the strange notion of a “cyclic time” in a strict sense, where R does not vary periodically in time but time itself moves, as it were, on a circle. This kind of cyclic time has been discussed by philosophers, but it has not found any use in science and is generally thought to be absurd (Whitrow 1980, 39–41).

Although Friedmann did not express either physical or philosophical preferences for a particular world model, he seems to have been fascinated by the possibility of a periodic or cyclic universe. In a semipopular book published in Russian in 1923, *The World as Space and Time*, he elaborated on the subject:

Cases are also possible when the radius of curvature changes periodically. The universe contracts into a point (into nothing) and then increases its radius from the point up to a certain value, then again diminishes its radius of curvature, transforms itself into a point, etc. This brings to mind what Hindu mythology has to say about cycles of existence, and it also becomes possible to speak about “the creation of the world from nothing”, but all this should at present be considered as curious facts which cannot be reliably supported by the inadequate astronomical material.²

As is well known, Friedmann’s seminal works of 1922–1924 were ignored by contemporary physicists and astronomers, who also failed to pay attention to his novel conception of a cyclic and closed universe. Indeed, his work is a prime example of what is known as a premature discovery (Hetherington 2002).

Three years after Friedmann’s paper in the *Zeitschrift für Physik*, there appeared in the same journal a lengthy article by the Hungarian physicist Cornelius Lanczos, one of the pioneers in the early phase of relativistic cosmology and also a contributor of some significance to the early development of quantum mechanics. This article too was ignored by contemporary physicists—during the 1920s it received only

²Friedmann (2000, 109). The book was translated into French in 1997 (Luminet 1997, 99–214) and into German three years later (Friedmann 2000), introduced and annotated by Georg Singer.

a single citation—and neither has it been noticed by historians of physics and cosmology. In spite of its lack of impact, it deserves attention.

Without citing Friedmann, Lanczos investigated a world model which was not only closed in space but also in time, with time being constructed as a periodic coordinate (Lanczos 1925). Although he did not refer to Friedmann, I consider it probable that he read the paper of 1922 and received some inspiration from it. As a theoretical cosmologist and a frequent contributor to the *Zeitschrift*, it is hard to believe that he failed to pay attention to Friedmann's work. For one thing, the paper of 1925 adopts a language close to that used by Friedmann. Not only did Lanczos speak of a "world period" (*Weltperiode*), the very term introduced by Friedmann, he also distinguished between space varying periodically in time and time being a cyclic parameter. He argued that the latter concept, which leads to an eternally recurring universe in the sense of Friedrich Nietzsche (whom he quoted), was contradictory and of no scientific use. Whatever his inspiration from Friedmann, Lanczos found for the world period the expression

$$T = \frac{4\pi^2 m}{h} R^2 \cong 10^{41} \text{ years} \quad (8.6)$$

where h is Planck's constant, R the Einstein radius for the static universe, and m the mass of an electron. He noted that the number was enormously greater than R/c , by a factor of about 10^{32} . However, the period T was not a period of an oscillating universe, for according to Lanczos his model was not really periodic in time. It is unclear to me how he understood his picture of what may appear to be a closed world varying in time with an enormous world period. The aim of his paper was not to suggest a new cosmological theory but to relate the quantum phenomena of the microcosmos to the structure of the macrocosmos, namely, to understand Planck's quantum of action in terms of cosmology. "The solution to the quantum riddles is hidden in the spatial and temporal closedness of the universe", he wrote Lanczos (1925, 80).

After having learned about Edwin Hubble's analysis of measurements of nebular redshifts, Einstein abandoned his previous insistence on the static cosmological solution and accepted the expanding universe as a superior alternative. In the spring of 1931, slightly before Lemaître published his idea of an exploding universe, Einstein belatedly recognized Friedmann's pioneering work which was "uninfluenced by observations" and decided that there no longer was any need for the cosmological constant. He discussed a Friedmann cyclic model, filled with pressureless matter ("dust") and containing no radiation but without extending it to possible previous or later cycles.³ From Friedmann's first equation, Einstein got the expression

³Steinhardt and Turok (2007, 177) suggest that Einstein's choice of investigating the periodic Friedmann solution was a result of his "philosophical predilections" and fascination of Spinoza's philosophy. The suggestion lacks documentary evidence as well as plausibility. See O'Raifeartaigh and McCann (2014) for a detailed analysis of Einstein's paper and an English translation of it.

$$\left(\frac{dP}{dt}\right)^2 = c^2 \frac{P_0 - P}{P} \quad (8.7)$$

where $P = P(t)$ is the curvature radius of the closed universe and P_0 its maximum value. He explained: “For small P (our idealization is invalid for the strict limit $P = 0$), P increases very rapidly. Then, as P increases, the speed of change dP/dt decreases ever more and becomes zero at the limiting value $P = P_0$, after which the entire process takes place in the opposite sense (i.e. with P decreasing at an increasing speed) (Einstein 1931, 237). Assuming that $P_0 - P$ was of the same magnitude as P_0 , he estimated from Hubble’s data that the current world radius was only about 100 million light years and the mean density of matter as high as 10^{-26} g/cm³. At the end of the paper, Einstein summarized what he considered to be its significance:

This theory is sufficiently simple that it can be conveniently compared with the astronomical data. It further shows how cautious one should be with large extrapolations of the time in astronomy. It is, first of all, remarkable that the general theory of relativity seems able to justify in a more natural way (namely, without the Λ term) Hubbel’s [sic] new data than the postulate of the quasi-static nature of space, which now has little empirical support.⁴

The 1931 paper is not among Einstein’s better known works. Yet it is noteworthy and that not only because it marked Einstein’s public abandonment of the cosmological constant but also because he explicitly formulated a version of what soon became known as the cosmological principle: “All places in the universe are equivalent [*gleichgültig*]”. Moreover, formally the model belonged to the big bang class; indeed it was the first model ever of this kind. Later the same year, Richard Tolman investigated Einstein’s model in greater detail, in particular by introducing thermodynamical considerations. By using relativistic thermodynamics, he obtained the surprising result that the expansion and contraction of the model universe were not accompanied by increase in entropy, from which he suggested that they “could presumably be repeated over and over again” (Tolman 1931b, 1761). Tolman also provided a general expression of the way in which the radius of the Einstein cyclic universe varied in time. His result was

$$\sqrt{\frac{R}{R_m}} = \sin\left(\frac{t}{R_m} + \sqrt{\frac{R}{R_m}\left(1 - \frac{R}{R_m}\right)}\right) \quad (8.8)$$

where R_m is a constant signifying the upper limit of R . The radius would expand from $R = 0$ at $t = 0$ to R_m at $t = \pi R_m/2$ and then return to zero at $t = \pi R_m$. Written in a parametric form (Tolman 1934, 413), the expression represents a cycloid in the Rt plane given by

⁴Hubble’s name was systematically misspelled as “Hubbel”.

$$R = \frac{\alpha}{3}(1 - \cos \Psi), \quad t = \frac{\pi}{6}(1 - \sin \Psi) \quad (8.9)$$

where α denotes the constant quantity $8\pi\rho R^3$. The radius will oscillate between $R = 0$ and $R_m = \alpha/3$ at $t = \pi\alpha/6$.

The cyclic or pulsating model Einstein proposed in 1931 held no special significance for him, such as shown by the model he developed the following year in collaboration with Willem de Sitter (Einstein and de Sitter 1932). In the well-known Einstein-de Sitter model, the pressure and the cosmological constant were assumed to be zero, as in the earlier model, but it also assumed a flat space and consequently was steadily expanding according to $R(t) \sim t^{2/3}$. In Einstein's view, the significance of his papers of 1931–1932 was not so much that they described new cosmological models but that they demonstrated that the cosmological constant was unnecessary. This was an “incomparable relief”, as he wrote to Tolman in the summer of 1931. Einstein further pointed out the difficulty with the singularities formally appearing at $t = 0$ and $t = t_{\max}$, suggesting that they might disappear in a more realistic version of the model. Tolman responded:

When I first saw your proposed quasi-periodic solution for the cosmological line element, I was very much troubled by the difficulties connected with the behaviour of the model in the neighborhood of the points of zero proper volume. The remarks in your letter, however, pointing out that the actual inhomogeneity in the distribution of matter might make the idealized treatment fail in that neighborhood, seem to me very important... I think that it is pertinent to remark that from a physical point of view contraction to a very small volume could only be followed by renewed expansion. Hence all in all I am feeling much more comfortable about this difficulty, and indeed have just sent an article to the *Physical Review* discussing among other things the application of relativistic thermodynamics to quasi-periodic models of the universe.⁵

At the end of his paper of 1931, Einstein noted the time-scale difficulty; the much discussed problem that the time allowed by the cosmological model—he stated it to be about 10 billion years—was much smaller than the age of the stars and galaxies as estimated at the time (Kragh 1996, 73–79). This was a serious problem in the Einstein-de Sitter model, where $t = 2/3T_0$ with T_0 the Hubble time, which in the 1930s was believed to be about 1.8 billion years; and it was even more serious in the denser oscillating models where the present age must be less than 1.2 billion years. Einstein suggested that “one can try to get out of the difficulty by pointing out that the inhomogeneity of stellar matter makes our approximate treatment illusory” (Einstein 1931, 237). In a later survey of the cosmological problem, first published in 1945, he repeated his suggestion that the theory was “inadequate for very high density of matter” (Einstein 1953, 124). Although Howard Percy Robertson at Princeton University did not endorse Einstein's model, he agreed

⁵Tolman to Einstein, 14 September 1931, a response to Einstein to Tolman, 27 June 1931. Courtesy the Einstein Archives and Princeton University Press. Tolman's reference to his forthcoming paper was to Tolman (1931b).

that it was “emotionally more satisfactory” to assume that the field equations break down near $R = 0$ and leave room for a non-singular bounce (Robertson 1932, 224).

Unknown to Einstein, a physicist from Japan had investigated cyclic models of the universe a little earlier than himself. In September 1930, the Japanese theoretical physicist Tokio Takeuchi read a paper to the Physico-Mathematical Society of Japan on the cyclic universe which was published the following year (Takeuchi 1931).⁶ Apparently unaware of Friedmann’s earlier work, Takeuchi constructed a complicated cyclic line element which he claimed was “in agreement with the view of Boltzmann”.⁷ From a philosophical point of view, he found a monotonically increasing universe to be “not pleasing”. His theory had the advantage not only of securing the eternity of the universe but also of avoiding singularities where the energy-momentum tensor becomes infinite. For the total volume of the oscillating universe, he found the expression

$$V(t) = e^{3/2 \sin kt} 2\pi^2 R^3 \quad (8.10)$$

where k and R are constants and the velocity of light is taken as unity. The universe will thus reach a maximum size at $kt = \pi/2$. Inspired by Tolman, he discussed the thermodynamical properties of his cyclic model universe, including its brightness and the transformation of matter into energy. Published in a not widely known Japanese journal, Takeuchi’s theory attracted almost no attention. It was however noticed by Tolman, who dismissed it as artificial and devoid of physical interest (Tolman 1931b, 1764).

8.3 A Controversial Universe

The time-scale problem was not only a concern of Einstein’s it also worried Willem de Sitter who for a time thought that it justified a kind of pulsating universe, although not in Einstein’s sense. The Dutch astronomer speculated that the universe may have “shrunk during an infinite time from an infinite radius to a minimum value, . . . increasing again afterwards, the minimum being reached a few thousand million years ago” (De Sitter 1931a, 7).⁸ However, he realized that

⁶Takeuchi wrote several papers on relativity, quantum theory, and cosmology in the years about 1930, some of them in the proceedings of the Physico-Mathematical Society and others in the *Zeitschrift für Physik*. For example, he investigated the hypothesis of a decreasing velocity of light within the framework of evolutionary cosmology, concluding that the decrease was only about 1 cm/sec/year (Takeuchi 1930).

⁷In fact, Boltzmann never advocated or discussed an oscillating universe. In 1895 he developed a remarkable scenario of a kind of multiverse, including “worlds” with a reversed entropic order, but he did not consider a series of such worlds changing periodically in time (Boltzmann 1895).

⁸A similar speculation appeared in (Robertson 1932, 224), who suggested that the universe “was originally shrinking and, having reached a finite lower limit, began to expand”.

there was no very good reason to advocate such a cosmic scenario. In fact, in a systematic study of cosmological solutions from the summer of 1931, he found that all oscillating models were ruled out as incompatible with empirical data (De Sitter 1931b). A somewhat similar critique came from Robertson in his influential review of cosmological models in the *Reviews of Modern Physics* from 1933. Robertson pointed out that Einstein's pulsating model required an unrealistically high value of the matter density, namely, about 10^{-27} g/cm³, and that this was several thousand times more than indicated by observations (Robertson 1933, 78). His own favourite cosmological model was the Lemaître-Eddington solution, where the universe expands gently and monotonically from an Einstein state.

In a careful study of the world models of Friedmann and Lemaître, Gawrilow Raschko Zaycoff, a Bulgarian physicist and physics teacher, commented on the repeated "births" and "deaths" of the oscillating universe. An examination of the possibility of a lower limit $R > 0$ led him to conclude that, irrespective the value of the cosmological constant, "there exist no periodic solutions to the gravitational equations of the cosmological problem" (Zaycoff 1933, 135). However, he left open the possibility of non-singular bounces in the case of a modification of the field equations.

De Sitter continued to speculate that considerations of a possible state of the universe before $R = 0$ might solve the time-scale difficulty. At a meeting of the Royal Astronomical Society in 1933, he suggested that perhaps the universe had once contracted to a point, with all the galaxies passing simultaneously through it some 3–5 billion years ago. By assuming that the stars had survived this "very vigorous" critical event, their true ages could be much longer than the age of the universe as based on the recession of the galaxies (De Sitter 1933a, 184). De Sitter believed that observations favoured either a steadily expanding universe of the Einstein-de Sitter type or an oscillating universe. But he admitted that the latter alternative did not appeal to him: "Personally I have, like Eddington, a strong dislike to a periodic universe, but that is a purely personal idiosyncrasy, not based on any physical or astronomical data" (De Sitter 1933b, 630).

Eddington not only disliked Lemaître's hypothesis of an ever-expanding universe with an origin a finite time ago, he was equally opposed to the idea of a cyclic universe, whether in its classical or relativistic version. In agreement with the critique raised against the cyclic universe in the nineteenth century, he found this idea to contradict the law of entropy increase, a law he considered to be absolutely fundamental. "I am no Phoenix worshipper", he admitted in 1928. "I would feel more content that the Universe should accomplish some great scheme of evolution and, having achieved whatever may be achieved, lapse back into chaotic changelessness, than that its purpose should be banalized by continual repetition. I am an Evolutionist, not a Multiplicationist. It seems rather stupid to keep doing the same thing over and over again" (Eddington 1928, 86).

The change from a static to an evolving universe did not cause Eddington to change his view and neither did Tolman's revision of the thermodynamics of the universe which he chose to ignore. His dismissal of the cyclic universe was not primarily scientific but rather based in his religious and moral sentiments: "From a

moral standpoint the conception of a cyclic universe, continually running down and continually rejuvenating itself, seems to me wholly retrograde” (Eddington 1935, 59, my emphasis). Eddington was not the only one to feel in this way. In a lecture of 1940, the distinguished American astrophysicist Henry Norris Russell expressed his surprise of “the wide-spread desire to believe in some cyclical restoration of however great intervals” (Russell 1940, 27). With regard to this question, which he considered to be aesthetically rather than religiously based, he sided with Eddington.

Although short-lived and merely a more elaborate version of what Friedmann had shown earlier, Einstein’s pulsating model attracted considerable attention, both scientifically and among a broader audience. In May 1931 Einstein went to Oxford to receive an honorary doctorate and give a series of three Rhodes Memorial Lectures. The second of the lectures, delivered on May 16, dealt with the “cosmologic problem” and included a discussion of his as yet unpublished model of the cyclic universe without a cosmological constant.⁹ In a book titled *God and the Astronomers* the theologian William Ralph Inge, Dean of St. Paul’s, commented on Einstein’s view “that the ponderable matter of the universe alternately expands and contracts” from a religious perspective: “This, if it is Einstein’s settled view, is a revolutionary change, for it means a return to the old theory of cosmic cycles, which has long attracted me” (Inge 1934, 50). Contrary to most other theologians and Christian thinkers, Inge subscribed to the view that an eternally existing universe was in full agreement with Christian thought. Einstein’s model was not actually perpetually cyclic, but Inge evidently thought it was.

In an important paper of 1933, Georges Lemaître studied the “annihilation of space”, that is, the problematic singularity where “the radius of space may pass through zero”.¹⁰ In this connection he used as an example what he called “Einstein’s cycloidal universe”. However, he concluded that the model could not be corrected because it was unable to provide an age of the universe (the time since the initial singularity) longer than 1–2 billion years. Lemaître tended to regret the nonphysical nature of oscillating solutions “from a purely aesthetic point of view”, which he explained as follows: “Those solutions where the universe expands and contracts successively while periodically reducing itself to an atomic mass of the dimensions of the solar system, have an indisputable charm and make one think of the Phoenix of legend” (Lemaître 1997, 679); (Lambert 1999, 152–155). Many years later, in his presentation to the Solvay Congress of 1958, he returned to the Phoenix universe in which “any detail of the contraction period should have been destroyed”. He

⁹The lecture was not published, but a brief summary of it appeared in *Nature*, vol. 127, p. 790. The blackboard Einstein used, provided with his calculations of the cyclic universe, was kept and can be seen at the Museum of the History of Science in Oxford. The formulae on the blackboard correspond closely to those in his published paper. See O’Raifeartaigh and McCann (2014).

¹⁰Lemaître (1933) was reprinted in 1972 in *Pontifical Academiae Scientiarum Scripta Varia*, no. 36: 107–181. In 1997 an English translation by M. A. H. MacCallum appeared in *General Relativity and Gravitation* (Lemaître 1997).

argued that the new expansion would result in a mass of gas in a state of maximum entropy. Keeping to his old idea of an exploding primordial atom as the source of the universe, he said:

On the contrary, the distribution coming out from fresh matter would be a distribution of minimum entropy, i.e. a very unprobable distribution, very far from thermodynamic equilibrium. ... The only feature it has in common with a gas is that it is formed of a great number of individual "molecules," but they have not the Maxwellian distribution which is the real characteristic of a gas. It is better to describe such a state of matter as being corpuscular radiation travelling along in every direction. ... I do not see how a useful cosmology can be built by starting from the Phenix nucleon gas. (Lemaître 1958, 9)

It is sometimes stated that Lemaître was in favour of the Phoenix universe model, but this is definitely not the case. He discussed it briefly at a few occasions, without ever advocating it. Indeed, being a Catholic priest, it would have been most surprising if he had adopted such a picture of the universe, traditionally being associated with materialism and atheism.

During the 1930s, oscillating models of the type first described by Friedmann were well known and included in the reviews of Robertson, Heckmann, Zaycoff, McVittie, and others. They entered the scientific literature alongside other relativistic models but were rarely seen as particularly important or interesting. In fact, no physicist or astronomer in the period expressed strong commitment to the idea, whereas a few, such as Eddington and de Sitter, reacted emotionally against it.

8.4 Richard Tolman and Cosmic Entropy

A thorough investigation of cyclic models was first undertaken by Richard Chase Tolman, the eminent and versatile American physicist who since 1922 worked as professor at the California Institute of Technology (Caltech). Tolman was a pioneer in applying general relativity and thermodynamics to the universe at large. For example, as early as 1922, he applied chemical equilibrium theory to the hypothetical formation of helium from hydrogen in a static Einstein universe, concluding that the relative abundances of the two elements could not be explained in this way (Tolman 1922). In later works he derived expressions for the total energy and entropy of the universe, first in the static case and next for models of the evolving universe. Tolman was the first to express the second law of thermodynamics in a covariant formulation and to discuss the cosmological consequences of it (Tolman 1931a).¹¹ He originally examined an expanding radiation-filled universe, in which case he concluded that periodic solutions could not occur. However, he also emphasized that "we must not conclude therefrom that periodic solutions would be of no interest for the actual universe" (Tolman 1931a, 1660).

¹¹Relativistic thermodynamics started much earlier. For a critical survey of the early development, see Liu and Chuang (1992).

A main point of his investigations from the early 1930s was the demonstration that if the relativistic form of thermodynamics were applied to the universe, it would lead to results very different from those based on the classical thermodynamical reasoning of the late nineteenth century. Significantly, there would be no justification for either entropic creation or the heat death, that is, the creation of the world in a low entropic state or the end of it in a state of maximum entropy. In his important textbook of 1934, *Relativity, Thermodynamics and Cosmology*, he formulated this general conclusion as follows: “It would seem wisest, if we no longer dogmatically assert that the principles of thermodynamics necessarily require a universe which was created at a finite time in the past and which is fated for stagnation and death in the future” (Tolman 1934, 444).

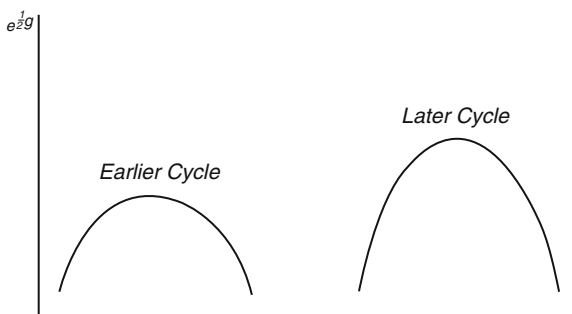
Einstein was well aware of Tolman’s works, which he admired. He corresponded with him in the 1930s, and in early 1931, when he spent a couple of months in Pasadena, he had intensive discussions with the American theorist. Einstein’s paper on the pulsating model, submitted to the Prussian Academy of Sciences after his return to Berlin, was probably indebted to his interactions with Tolman. Einstein showed his appreciation of Tolman’s contributions in a laudatory review he wrote of *Relativity, Thermodynamics and Cosmology* in 1934. In the review, which was published in *Science* and appeared in the original German (!), he explained his reasons for having abandoned the cosmological constant.¹²

According to Tolman, whereas a cyclic universe contradicted the classical version of the second law of thermodynamics, in relativity theory processes could take place without any increase in entropy at all. A world model such as Einstein’s periodic universe could expand and contract reversibly without increase in entropy, he concluded (Tolman 1931b). Moreover, extending his analysis to systems with irreversible processes he found that it was possible for a closed universe to undergo a continual series of cycles if only $\Lambda \leq 0$. In general, “relativistic thermodynamics could not impose restrictions which would prevent such a series of expansions and contractions” (Tolman 1932a, 331). According to Tolman, the simple picture of identical cycles had to be replaced with one in which each new cycle became greater than the previous cycle, both with respect to the period and the maximum value of the curvature radius. As to the entropy, although it would increase from one cycle to the next, it would never attain or approach a limit of maximum entropy. Alluding to the discussion in the nineteenth century—or perhaps to Eddington’s advocacy of the heat death scenario—he concluded that a succession of expansions and contractions could occur “without ever coming to that dreadful final state of quiescence predicted by the classical thermodynamics” (Tolman 1932b, 372) (Figure 8.1).

Much later Peter Landsberg and David Park examined the entropy in an oscillating universe model by means of computer experiments (Landsberg and

¹²*Science* 80 (1934): 358. A more visible result of Einstein’s stay in Pasadena was a brief paper he wrote jointly with Tolman and Boris Podolsky on the philosophical problems of quantum mechanics (Einstein et al. 1931).

Fig. 8.1 Tolman's illustration of two cycles of the oscillating universe, with the later cycle being greater than the earlier one. The quantity $e^{g/2}$ represents the radius of curvature (Tolman 1934, 443).



Park 1975).¹³ Their results were in broad agreement with those found by Tolman, including that successive cycles become larger and larger. They also found, again in agreement with Tolman, that the entropy increases continuously and can thus, even in an oscillating universe, be used as a measure for the direction of time. These results were confirmed by Landsberg and Reeves (1982), who further showed that the model collapses faster than it expands, i.e. $|dR/dt|_{final} > |dR/dt|_{initial}$.

Tolman's analysis of a pulsating universe filled with matter and radiation was valid for what he called quasi-periodic models, corresponding to an expansion of the universe from $R = 0$ to an upper limit and followed by a contraction to the same singular state. "Strictly periodic solutions" were outside the mathematical analysis, but Tolman thought that such a continual series of successive expansions and contractions might well be possible, indeed highly probable, from a physical point of view. He convinced himself that the conception of an ever-oscillating universe was not only "conceivable" but also "reasonable". According to Tolman, in a physically realistic universe, supposed to be not strictly isotropic and homogeneous, the singularity would not appear. Although he was unable to provide a plausible physical mechanism for the bounce that supposedly led to a new cycle, he suggested that such passages through $R = 0$ (or $R = R_{min}$) were "physically inevitably necessary" (Tolman 1934, 439). In lack of a mechanism, he suggested an analogy to the behaviour of an elastic ball bouncing up and down from the floor: Newton's second law in the form $d^2x/dt^2 = -g$ can describe how the ball rises to a maximum height and subsequently falls to the floor, but it cannot describe the mechanism of reversal when the ball hits the floor; considerations concerning the elastic properties of the ball and the floor have to be taken into account, and these are not provided by the equation of motion.

Tolman was well aware of the danger of confusing a cosmological model with the real universe, something he often warned against. Yet, in spite of his cautious

¹³All considerations of the entropy of the universe, whether supposed to be closed or not, rest on the assumption that the idea of entropy can be applied to the universe as a whole. It was and still is rarely realized that this is a questionable assumption. According to Robert Wald, there is "no reason to expect that there will be a meaningful notion of the 'total entropy of the universe' " (Wald 2006, 396).

attitude, he seems to have believed that his analysis justified an ever-existing oscillating universe rather than an explosive universe of Lemaître's type. He found it "difficult to escape the feeling that the time span for the phenomena of the universe might be most appropriately taken as extending from minus infinity in the past to plus infinity in the future" (Tolman 1934, 486). In an important paper co-authored by his student Morgan Ward, he showed from Einstein's field equations an early version of the singularity theorem, namely, that a contracting closed universe with $\Lambda = 0$ will end in a zero volume.¹⁴ He nonetheless believed that "from a physical point of view . . . we might expect contraction to the lower limit to be followed by a renewed expansion" and spoke in favour of the "possibility that the actual universe or parts thereof might also exhibit such a continued succession of expansions and contractions" (Tolman and Ward 1932, 837, 843). It is unclear if he realized that there can only be a limited number of preceding cycles, such as follows from the increase in entropy from cycle to cycle. If he did, he did not mention it.

Although Tolman certainly found oscillating models to be of great interest, his interest did not extend to an emotional or philosophical commitment. In early 1932, he gave an address to the Philosophical Union, a society at the University of California at Los Angeles, in which he discussed the new cosmological models and his own work on the thermodynamics of the universe. "In studying the problem of cosmology", he said, "we are immediately aware that the future fate of man is involved in the issue, and we must hence be particularly careful to keep our judgments uninfected by the demands of religion, and unswerved by human hopes and fears". He continued:

Thus, for example, what appears now to be the mathematical possibility for a highly idealized conceptual model, which would expand and contract without ever coming to a final state of rest, must not be mistaken for an assertion as to the properties of the actual universe, concerning which we still know all too little. . . . Although I believe it is appropriate to approach the problems of cosmology with feelings of awe for their vastness and of exultation for the temerity of the human spirit in attempting their solution, they must also be approached at the same time with the keen, balanced, critical and skeptical objectivity of the scientist. (Tolman 1932b, 373)

A similar statement appeared in his book of 1934, where he warned against a realistic interpretation not only of Lemaître's new big bang model but also of cyclic models (Tolman 1934, 488). Tolman's lack of strong commitment to an oscillating universe is confirmed by his later publications in cosmology, where this kind of model was in no way highlighted. For example, in a survey article of 1936, he merely mentioned the oscillating model as one possibility among others (Tolman 1937, 37).

The oscillating model was not only problematic from a theoretical point of view; it was also confronted by observational problems. For one thing, it shared with the Einstein-de Sitter open model the problem of an age of the universe that was shorter than the age of the Earth. In addition, it required a high average density of matter

¹⁴For an in-depth analysis of the history of singularity theorems, see Earman (1999).

to reverse the motion of the universe. In a closed universe with $\Lambda = 0$, the density must exceed the critical value given by $3H_0^2/8\pi G$, and with an accepted value of the Hubble constant about 500 km/s/Mpc, this meant a density greater than $10\text{--}28 \text{ g/cm}^3$. This was indeed a problem, but not a fatal one. In *The Realm of the Nebulae* of 1936, Hubble stated 10^{-30} g/cm^3 as a lower limit and 10^{-28} g/cm^3 as an upper limit. Other astronomers in the 1930s were willing to accept a mean density as high as 10^{-27} g/cm^3 . Given the state of uncertainty in the observations, a cyclic universe remained a possibility, if not a very likely one.

It should be noted that not all cyclic conceptions of the universe in the interwar period were based on relativity theory. The older conception of an eternally regenerating universe in the style of the nineteenth century continued to be discussed, quite independent of the oscillating models based on relativistic cosmology. Although these ideas were separated from mainstream cosmology and ignored by most physicists and astronomers, they enjoyed a good deal of public support and were advocated by a few scientists of distinction (Kragh 1995); (Jaki 1974, 342–345). Hypotheses of a continually recycling universe with an equilibrium between organization and dissipation processes (or processes consuming and producing entropy) were suggested by, among others, Emil Wiechert and Walther Nernst in Germany, Robert Millikan and William MacMillan in the United States, and Oliver Lodge in England. These ideas had in common that they postulated an eternal universe that was perpetually creative but without ever approaching a heat death. Accepting the traditional notion of space being flat and static, they were not cosmological models in the sense of relativity theory. They were recycling or regenerative world pictures, not cyclic in the sense of exhibiting a temporal periodicity over long spans of time.

8.5 The Oscillating Universe in the 1950s

Much of the development in cosmology during the first two decades following World War II was concerned with the controversy between evolutionary models and the rival steady-state theory that Fred Hoyle, Thomas Gold, and Hermann Bondi introduced in 1948 (Kragh 1996). A cyclic universe obviously contradicted the basis of the steady-state theory, the perfect cosmological principle, whereas it was compatible with the idea of an exploding universe such as developed by George Gamow and his collaborators Ralph Alpher and Robert Herman in the late 1940s. On the other hand, Gamow's research programme focused on the very early universe following the big bang, whereas the geometry and long-time behaviour of the universe were thought to be of less importance.

Gamow speculated that the expanding universe was presumably the result of an earlier contraction but thought that "there is no sense in speaking about that 'prehistoric state' of the universe, since indeed during the state of maximum compression . . . no information could have been left from the earlier time if there ever was one" (Gamow 1951, 406). Still, the bouncing model appealed to him, such

as he made clear three years later, when he concluded that such a model was “much more satisfactory” than the finite-age explosion model of Lemaître. His picture was this:

Thus we conclude that our Universe has existed for an eternity of time, that until about five billion years ago it was collapsing uniformly from a state of infinite rarefaction; that five billion years ago it arrived at a state of maximum compression in which the density may have been as great as that of the particles packed in the nucleus of an atom . . . , and that the Universe is now on the rebound, dispersing irreversibly toward a state of infinite rarefaction. (Gamow 1954, 63)

Realizing that this was a speculation, he cautiously added that “from the physical point of view we must forget entirely about the pre-collapse”. As to the possibility of a periodic universe with either a finite or an infinite number of cycles, he dismissed it as being incompatible with observational data: “There is no chance that the present expansion will ever stop or turn into a collapse”, he wrote (Gamow 1952, 42). His universe was infinite in size and would continue to expand in an infinity of time. Gamow repeated the verdict in a popular paper of 1956, but with the reservation that there might be large amounts of dark matter in intergalactic space, in which case the universe would be of the oscillating type (Gamow 1956).

Some of the scientists who advocated oscillating models in the 1950s and later were motivated by a wish to avoid the perplexing problem of an origin of the universe. They found a primordial state, whether in the form of a singularity or a primaeval atom in the sense of Lematre, to be unacceptable from both an observational and a philosophical point of view. For example, this was the opinion of the French astrophysicist Alexandre Dauvillier, professor at the Collège de France, who believed that any notion of a finite-age universe was metaphysical and anthropomorphic. Those who entertained such ideas entered “an unintelligible metaphysical terrain”, he wrote. “Not only is the hypothesis [of Lemaître] not justified by observations, but it is a priori inadmissible because of its metaphysical character. It implies a supernatural creation *ex nihilo*, which remains outside scientific thought” (Dauvillier 1963a, 76, 95). Dauvillier further claimed that the success of Lemaître’s hypothesis was to a large extent due to its exploitation by writers and scientists by a mystical or religious orientation.

As an alternative he advocated an infinity of cosmic cycles, which he thought could provide a framework for understanding the cosmic rays and the formation of the chemical elements (Dauvillier 1955, 1963a,b). However, although he was a staunch supporter of “la théorie des cycles cosmiques”, it was not in the cosmological sense of general relativity but in the older and more restricted sense of energetic cycles occurring endlessly in the universe. Strangely, he did not refer to cyclic models in the tradition of Friedmann, Einstein, and Tolman. His ideas were closer in spirit to the earlier ones of Nernst, Millikan, and MacMillan than to the oscillating models based on relativistic cosmology.

In the 1950s, relativistic models of the oscillating universe were considered in particular by William Bowen Bonnor in England and Herman Zanstra in the Netherlands. Bonnor, a specialist in general relativity originally trained in physical chemistry, investigated in a series of works the problem of galaxy formation and

its connection to cosmological theories. In this context he argued for an oscillating universe in which the inhomogeneities of the early universe, out of which seeds galaxies were formed, were fossils from the preceding contraction (Bonnor 1954, 1957). Admitting that there was no known physical mechanism that would reverse the contraction, he argued that “appropriate pressure changes will cause the model to change from contraction to expansion without passing through a singular state” (Bonnor 1954, 20). He realized of course the old objection against the eternally oscillating universe based on the second law of thermodynamics but thought that it scarcely deserved to be taken seriously. As he wrote in a popular book of 1960, “it has never been properly shown how the Second Law of Thermodynamics affects the universe as a whole” (Bonnor 1960, 10).

Herman Zanstra is well known as a distinguished astronomer and astrophysicist, but his role as a cosmologist has remained unacknowledged. A specialist in the physics of gaseous and planetary nebulae, he served as professor of astronomy at the University of Amsterdam, and in 1961 he received the Gold Medal of the Royal Astronomical Society, arguably the most prestigious prize of the astronomy community (Plaskett 1974). Four years earlier, in a little known paper in the proceedings of the Royal Dutch Academy of Science, he examined in great detail oscillating models of the universe (Zanstra 1957). Although the work attracted very little attention, it deserves a place in the history of cosmological thought (Figure 8.2).¹⁵

Building on the earlier works of Tolman, Zanstra concluded that the oscillating universe was allowed observationally if only Hubble’s old distance scale was

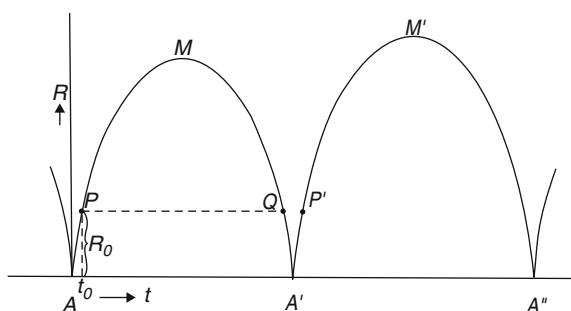


Fig. 8.2 The pulsating universe as depicted in (Zanstra 1957, 116). The cycle starts in A , and bounces at maximum density (A' , A'') are assumed possible. The state of the present universe is represented by P , at radius R_0 and age t_0 . Q marks the later contracting state. The second cycle $A'M'A''$ is greater than the first cycle AMA' .

¹⁵Zanstra (1957) appeared separately as circular no. 11 of the Astronomical Institute of the University of Amsterdam. It is not included in the ISI Web of Science and has not received attention by historians of science. The paper was abstracted in the *Astronomischer Jahresbericht* (vol. 57, 1957: 124–125) but not in the *Physics Abstracts*.

increased by a factor of 5 or more. Since observers at the 200-inch Hale telescope at Palomar Observatory had recently determined the Hubble time to be at least 5.4 billion years (Humason et al. 1956), as compared to the earlier estimated 1.8 billion years, he thought that an oscillating universe was a possibility. As to the question of whether the contraction of the universe could be reversed at a very small radius, or alternatively would end in a singularity, Zanstra concluded: “To stop the compression of the universe so that it can be followed by expansion would require a high negative pressure, which seems to be physically excluded” (Zanstra 1957, 114). Still, this result assumed the validity of the known laws of physics, general relativity included, at extremely high densities, and Zanstra was willing to question the assumption. In this regard he was in good company, for authorities such as Einstein and Tolman had advocated a somewhat similar line of thought (Einstein 1931, 235); (Tolman 1934, 438–439); (Earman 1999, 240–242).

Following Tolman’s analysis and assuming a repetition of cycles, Zanstra argued that with each new cycle, the universe would grow bigger, with greater values of the maximum curvature radius and the period. During the final phase of a contraction, near the minimum value of R , he found that the temperature would increase drastically and at $R = R_{min}$ the matter of the compressed universe would consist of a hot gas of electrons and protons. From this state of maximum compression, a new cycle would start afresh. However, based on thermodynamical arguments, he concluded that the oscillating universe could not have existed prior to a certain time, that is, it could only have been preceded by a finite number of earlier cycles: “Since at each reversal point a substantial more or less fixed amount of radiation is added, there cannot have been a whole pulsation prior to the reversal point where the radiation is less than this fixed amount” (p. 119). A similar result, based on the finite value of the entropy per baryon, became important in cosmology only about a decade later (Kragh 2009).

In spite of his sympathy for the oscillating universe, Zanstra found that it did not quite live up to what he called the “philosophical desires” of how the universe should evolve. These desires he formulated in three principles:

1. The universe must exist eternally, both in the past and in the future.
2. The universe must be self-regenerating and never end up in a state of thermal equilibrium.
3. Over long intervals of time, the universe must remain unchanged.

While ever-expanding models, such as of the Lemaître or Einstein-de Sitter type, violated all three of the principles, the oscillating model satisfied the second one: “Only one of the philosophical desires can be satisfied if reversal is assumed and none under ordinary laws” (p. 121).¹⁶ This made the cyclic universe appealing, but not appealing enough.

¹⁶It would seem that the steady-state universe, based on the perfect cosmological principle, satisfied all three principles, but Zanstra argued that it failed on account of the third principle.

Zanstra returned to the oscillating universe in a paper of 1967 in which he reconsidered the problem of the finite number of past cycles that was responsible for the violation of the first of his philosophical desires. He now introduced the highly unorthodox idea of “a series of occult non-physical interventions at every compression” in order to maintain a cyclic universe with an eternal existence in the past (Zanstra 1967, 39). These occult forces he described as originating in a conscious spiritual reality, a divine being of some sort. His attempt to introduce spiritual philosophy in science was politely ignored by his fellow astronomers.¹⁷

8.6 Negative Pressure as a Saving Device

While Zanstra dismissed a negative pressure as unphysical in his paper of 1957, in his later article in *Vistas in Astronomy*, he referred to the hypothesis such as introduced by William (McCrea 1951) and adapted to cyclic models by the Polish physicist Jaroslav (Pachner 1965). As early as 1934, Lemaître had pointed out that the vacuum corresponds to an ideal fluid with pressure and energy density given by

$$p = -\rho c^2 \quad \text{and} \quad \rho = \frac{\Lambda c^2}{8\pi G} \quad (8.11)$$

In the language of later cosmology, the equation of state for such a fluid is given by $w = -1$.¹⁸ In 1951 McCrea introduced the concept in a revised version of the steady-state theory but without assigning any direct physical effects to it (McCrea 1951; Kragh 1999). His conceptual innovation reappeared in a bouncing non-singular model that George McVittie proposed in 1952 and in which the universe contracted to a minimum value after which it would continually expand (McVittie 1952). However, it took several years until ideas of a negative pressure made an impact on cosmological thinking.

When negative pressure was incorporated into cosmology, it resulted in a proliferation of models (Harrison 1967; Clifton and Barrow 2007), and with the emergence of inflation theory in the early 1980s, the negative pressure associated with Λ became almost fashionable. Among the models that made use of the notion in order to avoid a singular state, there was a class of universes oscillating gently

¹⁷Zanstra gave an exposition of his metaphysical beliefs, which included telepathy and other parts of parapsychology, in a series of lectures at the philosophy department of the University of Michigan 1959–1960. The lectures were published as Zanstra (1962).

¹⁸Lemaître (1934). In cosmology the equation of state is given by a dimensionless number $w \equiv p/\rho$, where p is the pressure and ρ the energy density equal to the mass density times c^2 . For an ordinary gas, $w = 0$, whereas for radiation and relativistic matter, as in the early universe, $w = 1/3$. It can be shown that for any $w < -1/3$, the expansion of the universe is accelerating. One example is dark energy in the form of the cosmological constant, where $w = -1$. The even more extreme case $w < -1$ is characteristic of so-called phantom energy.

without bangs or crunches, of which Pachner's model of 1965 was an early example. However, I shall not here deal with these post-1960 cyclic models (see Kragh 2009) except pointing out that the idea of a negative pressure helped oscillating models to survive at a time when few cosmologists considered them to be a viable alternative to the emerging standard big bang picture.

The singularity theorems proved by Roger Penrose, Stephen Hawking, and others in the middle of the 1960s demonstrated that cosmic singularities were nearly unavoidable and therefore raised serious doubts as to the possibility of a bounce from one cycle to the next (see Earman (1999) and sources mentioned therein). On the other hand, the problem of cosmic singularities was not specifically related to oscillating models, and none of the investigations that led to the singularity theorems mentioned these models as particularly problematic. Moreover, workers in the field realized that their arguments for singularities were not waterproof. Thus, in 1956 Arthur Komar showed that cosmological singularities are to be expected under very general hypotheses, but he also noted that a negative pressure could prevent the occurrence of singular states (Komar 1956, 546). Hawking similarly referred to the negative energy C -field in the Hoyle-Narlikar steady-state theory as a possible way to avoid the cosmic singularity (Hawking 1966, 521).

With or without the hypothetical negative pressure, the oscillating universe faced serious problems of both an observational and theoretical kind. Nonetheless, this class of models continued to attract attention and be investigated by a minority of cosmologists, eventually leading to the revival of interest in the twenty-first century referred to in the introduction. But this is a story that goes beyond the limit of the present investigation.

Acknowledgements I thank the Caltech Institute Archives for permission to quote from Tolman's unpublished correspondence. Tilman Sauer kindly provided me with copies of letters between Einstein and Tolman, for which I am grateful.

References

- Boltzmann, L. (1895). On certain questions in the theory of gases. *Nature*, 51, 483–485.
- Bonnor, W. (1954). The stability of cosmological models. *Zeitschrift für Astrophysik*, 35, 10–20.
- Bonnor, W. (1957). La formation des nébuleuses en cosmologie relativiste. *Annales de l'Institut Henri Poincaré*, 15, 158–172.
- Bonnor, W. (1960). Relativistic theories of the universe. In H. Bondi et al. (Eds.), *Rival theories of cosmology* (pp. 1–11). London: Oxford University Press.
- Clifton, T. & Barrow, J. D. (2007). Ups and downs of cyclic universes. *Physical Review D*, 75, 043515.
- Dauvillier, A. (1955). *Cosmologie et chimie*. Paris: Presses Universitaires de France.
- Dauvillier, A. (1963a). *Les hypothèses cosmogoniques: Théories des cycles cosmiques et des planètes jumelles*. Paris: Masson et Cie.
- Dauvillier, A. (1963b). Les hypothèses cosmogoniques et la théorie des cycles cosmiques. *Scientia*, 98, 121–126.
- De Sitter, W. (1931a). The expanding universe. *Scientia*, 49, 1–10.

- De Sitter, W. (1931b). Some further computations regarding non-static universes. *Bulletin of the Astronomical Institutes of the Netherlands*, 6, 141–145.
- De Sitter, W. (1933a). Contribution to discussion. *Observatory*, 56, 184.
- De Sitter, W. (1933b). On the expanding universe and the time-scale. *Monthly Notices of the Royal Astronomical Society*, 93, 628–634.
- Earman, J. (1999). The Penrose-Hawking singularity theorems. In H. Goenner et al. (Eds.), *The expanding worlds of general relativity* (pp. 235–270). Boston: Birkhäuser.
- Eddington, A. S. (1928). *The nature of the physical world*. Cambridge: Cambridge University Press.
- Eddington, A. S. (1935). *New pathways in science*. Cambridge: Cambridge University Press.
- Einstein, A. (1931). Zum kosmologischen Problem der allgemeinen Relativitätstheorie. In *Sitzungsberichte der Preussischen Akademie der Wissenschaften* (pp. 235–237).
- Einstein, A. (1953). *The meaning of relativity*. Princeton: Princeton University Press.
- Einstein, A. & de Sitter, W. (1932). On the relation between the expansion and the mean density of the universe. *Proceedings of the National Academy of Sciences*, 18, 213–214.
- Einstein, A., Tolman, R. C. & Podolsky, B. (1931). Knowledge of past and future in quantum mechanics. *Physical Review*, 37, 780–781.
- Friedmann, A. (1922). Über die Krümmung des Raumes. *Zeitschrift für Physik*, 10, 377–386.
- Friedmann, A. (2000). *Die Welt als Raum und Zeit*. Frankfurt am Main: Harri Deutsch.
- Friedmann, A., & Lemaître, G. (1997). *Essais de cosmologie. Précédé de : Invention du Big Bang*. Ed. Jean-Pierre Luminet. Paris: Editions du Seuil.
- Gamow, G. (1951). The origin and evolution of the universe. *American Scientist*, 39, 393–407.
- Gamow, G. (1952). *The creation of the universe*. New York: Viking Press.
- Gamow, G. (1954). Modern cosmology. *Scientific American*, 190, 55–63.
- Gamow, G. (1956). The evolutionary universe. *Scientific American*, 192, 136–154.
- Harrison, E. R. (1967). Classification of uniform cosmological models. *Monthly Notices of the Royal Astronomical Society*, 137, 69–79.
- Hawking, S. W. (1966). The occurrence of singularities in cosmology. *Proceedings of the Royal Society of London A*, 294, 511–521.
- Hetherington, N. S. (2002). Theories of an expanding universe: Implications of their receptions for the concept of scientific prematurity. In E. B. Hook (Ed.) *Prematurity in scientific discovery: On resistance and neglect* (pp. 109–123). Berkeley: University of California Press.
- Humason, M. J., Mayall, N. U. & Sandage, A. R. (1956). Redshifts and magnitudes of extragalactic nebulae. *Astrophysical Journal*, 61, 97–162.
- Inge, W. R. (1934). *God and the astronomers*. London: Longmans, Green and Co.
- Jaki, S. L. (1974). *Science and creation: From eternal cycles to an oscillating universe*. Edinburgh: Scottish Academic Press.
- Komar, A. (1956). Necessity of singularities in the solution of the field equations of general relativity. *Physical Review*, 104, 544–546.
- Kragh, H. (1995). Cosmology between the wars: The Nernst-MacMillan alternative. *Journal for the History of Astronomy*, 26, 93–115.
- Kragh, H. (1996). *Cosmology and controversy: The historical development of two theories of the universe*. Princeton: Princeton University Press.
- Kragh, H. (1999). Steady-state cosmology and general relativity: Reconciliation or conflict? In H. Goenner et al. (Eds.), *The expanding world of general relativity* (pp. 377–402). Boston: Birkhäuser.
- Kragh, H. (2008). *Entropic creation: Religious contexts of thermodynamics and cosmology*. Aldershot: Ashgate.
- Kragh, H. (2009). Continual fascination: The oscillating universe in modern cosmology. *Science in Context*, 22, 587–612.
- Lambert, D. (1999). *Un atome d'univers: La vie et l'oeuvre de Georges Lemaître*. Brussels: Racine.
- Lanczos, C. (1925). Über eine zeitlich periodische Welt und eine neue Behandlung des Problems der Ätherstrahlung. *Zeitschrift für Physik*, 32, 56–80.

- Landsberg, P. T., & Park, D. (1975). Entropy in an oscillating universe. *Proceedings of the Royal Society of London A*, 346, 485–495.
- Landsberg, P. T., & Reeves, G. A. (1982). Heat death and oscillation in model universes containing interacting matter and radiation. *Astrophysical Journal*, 262, 432–441.
- Lang, K., & Gingerich, O. (Eds.). (1979). *A source book in astronomy and astrophysics 1900–1975*. Cambridge, MA: Harvard University Press.
- Lemaître, G. (1933). L'univers en expansion. *Annales de Sociétés Scientifiques de Bruxelles*, 53, 51–85.
- Lemaître, G. (1934). Evolution of the expanding universe. *Proceedings of the National Academy of Sciences*, 20, 12–17.
- Lemaître, G. (1958). The primeval atom hypothesis and the problem of the clusters of galaxies. In R. Stoops (Ed.) *La structure et l'évolution de l'univers* (pp. 1–32). Brussels: Coudenberg.
- Lemaître, G. (1997). The expanding universe. *General Relativity and Gravitation*, 29, 641–680.
- Liu, C. (1992). Einstein and relativistic thermodynamics in 1952: A historical and critical study of a strange episode in the history of modern physics. *British Journal for the History of Science*, 25, 185–206.
- McCrea, W. H. (1951). Relativity theory and the creation of matter. *Proceedings of the Royal Society A*, 206, 562–575.
- McVittie, G. C. (1952). A model universe admitting the interchangeability of stress and matter. *Proceedings of the Royal Society A*, 211, 295–301.
- O'Raiheartaigh, C., & McCann, B. (2014). Einstein's cosmic model of 1931 revisited: An analysis and translation of a forgotten model of the universe. *European Physical Journal H*, 39, 63–86.
- Pachner, J. (1965). An oscillating isotropic universe without singularity. *Monthly Notices of the Royal Astronomical Society*, 131, 173–176.
- Plaskett, H. H. (1974). Herman Zanstra. *Quarterly Journal of the Royal Astronomical Society*, 15, 57–64.
- Robertson, H. P. (1932). The expanding universe. *Science*, 76, 221–226.
- Robertson, H. P. (1933). Relativistic cosmology. *Reviews of Modern Physics*, 5, 62–90.
- Russell, H. N. (1940). The time-scale of the universe. *Science*, 92, 19–27.
- Steinhardt, P. J., & Turok, N. (2002). A cyclic model of the universe. *Science*, 296, 1436–1439.
- Steinhardt, P. J., & Turok, N. (2007). *Endless universe: Beyond the big bang*. New York: Doubleday.
- Takeuchi, T. (1930). Über die Abnahme der Lichtgeschwindigkeit. *Zeitschrift für Physik*, 69, 857–859.
- Takeuchi, T. (1931). On the cyclic universe. *Proceedings of the Physico-Mathematical Society of Japan*, 13, 166–177.
- Tolman, R. C. (1922). Thermodynamic treatment of the possible formation of helium from hydrogen. *Journal of the American Chemical Society*, 44, 1902–1908.
- Tolman, R. C. (1931a). On the problem of the entropy of the universe as a whole. *Physical Review*, 37, 1639–1660.
- Tolman, R. C. (1931b). On the theoretical requirements for a periodic behaviour of the universe. *Physical Review*, 38, 1758–1771.
- Tolman, R. C. (1932a). Possibilities in relativistic thermodynamics for irreversible processes without exhaustion of free energy. *Physical Review*, 39, 320–336.
- Tolman, R. C. (1932b). Models of the physical universe. *Science*, 75, 367–373.
- Tolman, R. C. (1934). *Relativity, thermodynamics and cosmology*. Oxford: Oxford University Press.
- Tolman, R. C. (1937). The present status of cosmology, II. *Scientific Monthly*, 44, 20–40.
- Tolman, R. C., & Ward, M. (1932). On the behavior of non-static models of the universe when the cosmological term is omitted. *Physical Review*, 39, 835–843.
- Troop, E. A., Frenkel, V. Ya., & Chernin, A. D. (1993). *Alexander A. Friedmann: The man who made the universe expand*. Cambridge: Cambridge University Press.
- Wald, R. M. (2006). The arrow of time and the initial conditions of the universe. *Studies in History and Philosophy of Modern Physics*, 37, 394–398.

- Weinberg, S. (1977). *The first three minutes: A modern view of the origin of the universe*. London: Trinity Press.
- Whitrow, G. J. (1980). *The natural philosophy of time*. Oxford: Oxford University Press.
- Zanstra, H. (1957). On the pulsating or expanding universe and its thermodynamical aspect. *Proceedings, Koninklijke Nederlandse Akademie van Wetenschappen, Series B*, 60, 285–307.
- Zanstra, H. (1962). *The construction of reality: Lectures on the philosophy of science, theory of knowledge, and the relation between body, mind, and personality*. Oxford: Pergamon Press.
- Zanstra, H. (1967). Thermodynamics, statistical mechanics and the universe. *Vistas in Astronomy*, 10, 23–43.
- Zaycoff, R. (1933). Zur relativistischen Kosmogonie. *Zeitschrift für Astrophysik*, 6, 128–137, 193–197.

Chapter 9

Inflation and the Origins of Structure



Chris Smeenk

9.1 Introduction

The initial motivations for a physical theory are sometimes rendered dubious or superfluous by later work. The epistemic load borne by motivating ideas in the first stage of theoretical construction is shifted onto other ideas as work proceeds, leaving the original arguments with a largely ornamental rather than structural role. For example, Einstein described one of the three foundational ideas of general relativity (GR) as Mach's principle, roughly speaking the claim that spacetime geometry should be fully determined by the distribution of matter without appeal to "background structures." This principle was one of Einstein's guiding ideas in the discovery of GR, but few modern relativists grant it the same pride of place in understanding the foundations of the theory he created.¹ Contemporary arguments in favor of accepting GR as the best available classical theory of gravitation barely mention Mach's principle. The contemporary path to justification of our scientific theories often does not recapitulate the path to discovery.

The central idea of inflationary cosmology is that the early universe passed through a phase of exponential expansion driven by a scalar field displaced from the true minimum of its potential energy. Guth (1981) provided a rationale for this idea that proved to be quite persuasive: inflation nearly eliminates the need for special initial conditions required by the standard model of cosmology. It was soon discovered that inflation also suggested a solution to a long-standing problem in

¹The meaning of Mach's principle and its status has been a focal point for foundational discussions since Einstein's day; see Barbour and Pfister (1995) for an entry point to the recent literature.

C. Smeenk (✉)
University of Western Ontario, London, ON, Canada
e-mail: csmeenk2@uwo.ca

relativistic cosmology: what is the origin of the seeds for the formation of structure in the universe? A recent textbook draws a distinction between the original rationale for inflation, as a “theory of initial conditions,” and a rationale based on predictions for the seeds of structure, as a “theory of the origins of structure”:

... [T]hese problems [related to initial conditions] can no longer be regarded as the strongest motivation for inflationary cosmology because it is not at all clear that they could ever be used to falsify inflation. [...] By contrast to inflation as a theory of initial conditions, the model of inflation as a possible origin of structure in the Universe is a powerfully predictive one. Different inflation models typically lead to different predictions for observed structures, and observations can discriminate strongly between them. ... *Inflation as the origin of structure is therefore very much a proper science of prediction and observation.* (Liddle and Lyth 2000, 5; my emphasis)

Liddle and Lyth clearly regard the early motivations for inflation as not sufficiently empirical, unlike the case for inflation that can be made given its connection with structure formation. Below I will argue, in agreement with Liddle and Lyth, that there is an important contrast between the historical motivations for inflationary cosmology and the strongest case that can now be made in its favor. But I will suggest a different way of characterizing the contrast, based on how informative different bodies of data are regarding inflation.

The main strategy pursued by inflationary cosmologists is to treat various properties of the early universe as the consequences of the dynamical evolution of a scalar field (or fields) in the early universe. In his persuasive case for inflation, Guth (1981) emphasized that such evolution could lead to a uniform, flat universe for a large range of initial conditions. Shortly after Guth’s paper appeared, several groups of cosmologists formulated accounts of the creation of seeds for structure formation during inflation. The mechanism for generating density perturbations is the most fruitful consequence of inflation, in two different senses. First, the problems Guth emphasized in presenting the theory were regarded as “enigmas” of the standard model of cosmology, when they were discussed at all. By way of contrast, the status of initial “seed” fluctuations was a major problem facing an appealing account of the origin of structure. Given that gravity should be the dominant force at large length scales, it is natural to suppose that structures such as galaxies evolved by the growth of small perturbations to an almost uniform initial distribution of matter. As described below in Section 9.2, while this gravitational instability picture was appealing, it seemed to require an extremely implausible initial distribution of matter. Inflation countered this objection and provided theorists with a way of calculating the density perturbations as a consequence of a stage of exponential expansion. Section 9.3 recounts the historical route by which cosmologists developed this account and contrasts the case of structure formation with the initial motivations for inflation.

Section 9.4 turns to the second sense in which this aspect of inflation has been particularly fruitful, namely, in providing the grounds for a detailed comparison with an alternative approach to understanding the seeds for structure formation. Structure formation via topological defects was studied extensively as an alternative to inflation throughout the 1980s and 1990s. Observations of temperature fluctuations

in the cosmic microwave background radiation (CMBR) were able to discriminate between these two approaches and clearly favored inflation. To what extent does this success reflect that inflation has correctly identified the physics of the early universe, as opposed to exhibiting sufficient flexibility to accommodate the observations? The final concluding section attempts to move beyond the way in which questions regarding the empirical status of inflation have been couched in the physics literature, in terms of Popperian “falsifiability.”

9.2 Structure Formation in the Standard Model

By the early 1970s, two aspects of what (Weinberg 1972) dubbed the standard model of cosmology were well understood theoretically, supported by observational evidence and accepted as a starting point for further research by most cosmologists.² First, the expanding universe models of general relativity, the Friedmann-Lemaître-Robertson-Walker (FLRW) models, were taken to provide an approximate description of the overall structure and evolution of the universe at some suitably large length scale. In the early days of the field, cosmologists focused on these models for pragmatic reasons. Due to the symmetry assumed to hold in these models, the dynamics of general relativity is reduced to simple equations relating the scale factor $R(t)$ to the matter-energy distribution. Evidence accumulated that this symmetry was not just a useful simplification. In particular, the uniformity of the cosmic microwave background radiation (CMBR), first observed in 1965, supported the applicability of these models even in the early universe. By 1970, almost all cosmologists had accepted the FLRW models as a useful approximation and had turned to the more specific task of measuring the expansion with sufficient accuracy to choose the best model (see, e.g., Sandage 1970).

Second, the theory accounted for two striking features of the universe as relics of the “primeval fireball.” Nuclear reactions in the early universe governed by the rate of expansion leave a telling trace—a helium abundance of about 26–28% according to a calculation by Peebles (1967), in agreement with observations. Further development of the theory of big bang nucleosynthesis clarified the dependence of the primordial element abundances on various cosmological parameters. The CMBR was a second natural consequence of a hot big bang. In the early universe, radiation and matter are coupled due to interactions, but as the temperature drops low enough for the existence of stable nuclei, the universe becomes effectively transparent to photons.³ The photons then cool adiabatically with the expansion of the universe

²The steady-state theory was no longer a serious rival to the standard “big bang” model by this time, although a small group of proponents (including Hoyle, Narlikar, and others) continued to explore the idea and to challenge the empirical underpinnings of the big bang model (see Kragh 1996).

³Recombination refers to the process by which nuclei capture free electrons and form neutral hydrogen and helium (although “re-” is misleading, as there was no earlier time, in the standard

while maintaining a black-body spectrum, and they carry a tremendous amount of information regarding the universe at the time of recombination. Since 1965 a series of increasingly sophisticated observational missions have succeeded in extracting more and more of this information. Although subsequent research has enriched both ideas considerably, the fundamentals were in place by the early 1970s and are presented in the influential texts by Weinberg (1972) and Peebles (1971).

By contrast with these successes, the standard model lacked a compelling account of structure formation. Weinberg prefaced his discussion with the caveat that:

... [w]e still do not have even a tentative quantitative theory of the formation of galaxies, anywhere near so complete and plausible as our theories of the origin of the cosmic abundance of helium or the microwave background. (Weinberg 1972, 562)

Unlike the successful aspects of the standard model, in the case of structure formation, it has been much more difficult to link tractable pieces of theory to observations. This reflects the intrinsic difficulty of the subject, which requires integrating a broader array of physical ideas than required for the study of nucleosynthesis or the FLRW models. This section will give a brief overview of the development of the field up to 1980, focusing on the status of initial conditions for structure formation.

The ideas Weinberg (1972) described as a speculative part of the standard model were first explored by Lemaître (1933). Newtonian gravity enhances clumping of a nearly uniform distribution of matter. In the early stages of clumping, small fluctuations in density can be treated as first-order perturbations to a background cosmological model. This will be the case if the density contrast $\Delta =: \frac{\delta\rho}{\rho}$ is less than 1, where $\delta\rho$ is the density enhancement over the background density ρ . A theory of the evolution of small fluctuations must be supplemented on both ends, so to speak. The theory assumes as given an initial spectrum of small fluctuations that are then enhanced via dynamical evolution. An appealing possibility is that the dynamics is unstable, leading to exponential growth of small fluctuations. Then, like the onset of turbulence in fluid mechanics, details regarding the initial state would be relatively unimportant. On the other end, the theory extends up to the point when the fluctuations “freeze out” from the cosmological expansion and begin to collapse into structures with much higher density contrasts (such as $\Delta \approx 10^6$ for a typical galaxy). Developing a theory governing this later stage of structure formation poses enormous challenges: perturbation theory does not apply, and the non-gravitational interactions of the constituents of the collapsing region can no longer be ignored.⁴ Despite these limitations, the theory of structure formation via gravitational enhancement of nonuniformities covers a large dynamical range. If

model, at which stable nuclei existed). During recombination, photons decouple from matter as the cross section for Thomson scattering drops to zero.

⁴Modern studies of the nonlinear regime employ numerical simulations, although there are a number of analytic techniques that were developed to study nonlinear evolution during this time (e.g., Press and Schechter, 1974). See, e.g., Chapter 17 of Peacock (1999) for an introduction.

successful, it would provide a link between the physical processes in the very early universe responsible for the initial fluctuations and the observationally accessible imprints of perturbations at later times.

Lifshitz (1946) was the first to treat the evolution of linear perturbations to a background model in general relativity, only to reject gravitational instability as a viable account of structure formation. He showed that in an FLRW model, the density contrast as a function of time grows very slowly. This result is surprising given the contrast with the account of instability due to Jeans (1902). Jeans derived an equation governing the evolution of small perturbations of a fluid including Newtonian gravity and showed that the behavior of different modes depends on how their wavelength compares to a critical wavelength, the Jeans length λ_J .⁵ For modes with $\lambda = \lambda_J$, there is a balance between the pressure of the fluid, resisting collapse, and the gravitational force; perturbation modes with $\lambda < \lambda_J$ exhibit oscillatory behavior, whereas those with $\lambda > \lambda_J$ are unstable and grow exponentially. Physically, in the final case, the matter density is sufficient to trigger gravitational collapse, leading to exponential growth of the amplitude of the fluctuation. If such rapid growth occurred in an expanding background as well, it would be possible for galaxies to form via gravitational enhancement of thermal fluctuations in the matter density, which Lifshitz (and many others) took to be a reasonable posit for the initial conditions. In this case the fluctuations away from uniformity would be given by the Poisson distribution, $\Delta \propto N^{-1/2}$ for N particles; for a galaxy-scale lump of particles, say 10^{68} particles, thermal fluctuations would give a low-density contrast $\Delta_i \propto 10^{-34}$. However, Lifshitz showed that cosmological expansion works against gravitational instability, with the density contrast growing slowly ($\Delta(t) \propto t^{2/3}$) during the matter-dominated era in an expanding model. (Pressure prevents growth of the density contrast during the earlier radiation-dominated era.)⁶ If an initial fluctuation spectrum is imprinted at, say, $t_i = 1$ second (Bonnor 1956), time is too short for the fluctuations to grow into galaxies—with growth on the order of 10^{12} rather than the 10^{40} that is needed. Lifshitz concluded that gravitational instability fails to account for the formation of galaxies. Subsequent work on linear perturbation theory corrected and augmented Lifshitz's analysis in several significant respects but with little impact on this line of argument.⁷

In the 1950s and early 1960s, many theorists found this criticism so compelling that they pursued alternative accounts of structure formation. Gamow, for example, turned to developing a theory based on primeval turbulence.⁸ Lifshitz's line of

⁵See, e.g., Longair (2007, Chapter 11) or Weinberg (2008, Chapter 5), for modern introductions to linear perturbation theory.

⁶Lifshitz (1946) analyzed the behavior of small perturbations for two different equations of state, corresponding to radiation-dominated expansion, i.e., $p = \rho/3$, where p is the pressure and ρ the energy density, and matter-dominated expansion with $p = 0$. See, e.g., Longair (2007) for a modern treatment.

⁷See Peebles (1980, 20–25) and Longair (2006, Chapter 15) for historical overviews.

⁸Gamow and Teller (1939) advocated an account of structure formation based on gravitational instability that is undermined by Lifshitz's results (as Lifshitz explicitly noted). Gamow (1952)

argument reflects of an assessment of the plausibility of fluctuations at early times. Even a spectrum of thermal fluctuations is not immediately ruled out; Bonnor’s argument shows that thermal fluctuations at $t_i = 1s$ will not undergo sufficient growth, but one can treat t_i as a free variable and simply impose the fluctuation spectrum at an earlier time. Such an initial fluctuation spectrum is still mysterious, as we will see in more detail shortly.

The enigmatic nature of the initial conditions was not a sufficient objection to cosmologists who explicitly adopted a more phenomenological approach to galaxy formation (see, e.g., Harrison 1968, Peebles 1968 and Zeldovich 1965). All the available cosmological theories required some specification of the initial conditions, and the gravitational instability account is not obviously more objectionable in this respect. Furthermore, projecting backward to find the required initial conditions could provide insight into new physics relevant in the early universe. The discovery of the CMBR provided an important new constraint along with the potential to establish observationally the fluctuation spectrum at the time of decoupling. The phenomenological approach focused on giving a more precise characterization of the initial fluctuations that were required for gravitational instability along with a detailed account of their dynamical evolution over time. Throughout the 1970s, theorists developed competing accounts of structure formation with the common aim of describing the evolution of the different physical degrees of freedom involved—radiation, baryonic matter, and the gravitational field. Solving the complete set of equations capturing all of the details of their interactions and dynamics, the coupled Boltzmann-Einstein equations, would have been computationally intractable. But given the background of an FLRW model, different physical effects are dominant at different stages of evolution. Initial matter and radiation perturbations would in general be a combination of two distinct modes:⁹

- *adiabatic*: Fluctuations in energy density of nonrelativistic matter ρ_m matched by radiation fluctuations (also called “entropy perturbations”), $\frac{4}{3} \frac{\delta\rho_m}{\rho_m} = \frac{\delta\rho_r}{\rho_r}$,
- *isothermal*: Radiation is uniformly distributed, $\frac{\delta\rho_r}{\rho_r} = 0$, although the matter is nonuniformly distributed.

and Gamow (1954) are the original papers on the turbulence theory; see Peebles (1971) for a critical review of Gamow’s proposal and other similar ideas. Two other problems with the gravitational instability account were also important in motivating the search for alternatives. First, there is no preferred length or mass scale in general relativity (with the cosmological constant set to zero), so it is unclear how to introduce scales such as the mass of a typical galaxy (see Harrison (1967a) and Harrison (1967b) for a detailed discussion of this point). Second, alternative accounts often claimed to give natural explanations of features of galaxies, such as their rotation and spiral structure.

⁹This terminology is due to Zel’dovich and his collaborators. The factor of $\frac{4}{3}$ arises since the energy density of radiation is $\propto T^4$, compared to T^3 for matter (where T is the temperature). These are called “adiabatic” perturbations since the local energy density of the matter relative to the entropy density is fixed. A third mode—tensor perturbations, representing primordial gravitational waves—was not usually included in discussions of structure formation, since they do not couple to energy-density perturbations.

One can then analyze the evolution of these distinct perturbation modes through different stages of the universe's history. Prior to recombination, radiation ionizes the baryons and the photons, and free electrons are coupled via Thomson scattering. As a result, fluctuations in the baryonic matter and radiation move together like a single fluid (Peebles 1965); galactic-scale perturbations undergo acoustic oscillations during this phase. In the later matter-dominated era, radiation and matter decouple, and the matter fluctuations can be treated in isolation along the lines of Lifshitz's analysis, and galactic-scale perturbations grow with $\Delta(t) \propto t^{2/3}$.

There were also debates regarding the appropriate initial spectrum and later stages of structure formation (see Longair 2006). Two different schools of thought dominated the field: Zel'dovich's school focused on solutions in which large "blinis" (pancakes) formed first from adiabatic perturbations, fragmenting into galaxies and structures much later due to non-gravitational processes. The other school of thought led by Peebles developed a "bottom-up" scenario, in which initial isothermal fluctuations developed into protogalaxies with larger structures forming later by hierarchical clustering. Despite the stark differences between the accounts these theories gave of later stages of structure formation, they had similar implications for the epoch of recombination.

Both schools of thought also needed to address the evolution of density fluctuations, and there was a natural choice for the initial spectrum. Harrison (1970), Peebles and Yu (1970) and Zel'dovich (1972) proposed a scale-invariant (HPZ) spectrum, meaning that $\Delta|\lambda = \text{constant}$ when λ , the perturbations' wavelength, is equal to the Hubble radius, $\lambda = H^{-1}$.¹⁰ This spectrum lacks any characteristic length scale. For different wavelengths, the perturbation amplitude is fixed at different times: in an expanding universe, the wavelength λ increases with the scale factor $R(t)$, whereas the Hubble radius increases at a slower rate as the expansion slows.¹¹ (The Hubble radius is a length scale set by the rate of expansion.)¹² The Hubble radius "crosses" various perturbation wavelengths in an expanding model; a scale-invariant spectrum deserves the name since the perturbations have the same magnitude as the Hubble radius sweeps across different length scales. Estimates of the magnitude of density perturbations when length scales associated with galaxies

¹⁰In general, for a scale-invariant power spectrum, the Fourier components of the perturbations obey a power law, $|\delta_k|^2 \propto k^n$; the Harrison-Peebles-Zel'dovich spectrum corresponds to a choice of $n = 1$ (given that the volume element in the inverse Fourier transform is $\frac{dk}{k}$; for the other conventional choice, $k^2 dk$, we then have $n = -3$). The Hubble radius has the appropriate dimension, length: restoring c , it is given by $\frac{c}{H}$, and the Hubble constant H has units of km per second per megaparsec.

¹¹Since the perturbations grow with time, at a "constant time" the shorter wavelength perturbations have greater amplitudes for this spectrum. The difficulty with defining the spectrum of density perturbations in terms of "amplitude at a given time" is that it depends on how one chooses the constant time hypersurfaces.

¹²It is defined as $R_H = H^{-1}$, where H is the Hubble "constant" ($H = \frac{\dot{R}}{R}$); it is also called the speed of light sphere, given that objects moving with the expansion, at a distance R_H , appear to move at speed c .

cross the Hubble radius fall within the range $\Delta \approx 10^{-3}$ – 10^{-4} . In addition, the initial perturbations were often also assumed to be “random” in the sense that the mass found within a sphere of fixed radius has a Gaussian distribution (for different locations of the sphere).

Two features of HPZ spectrum are particularly puzzling. The first puzzle arises from the causal structure of the FLRW models. Even though the distance between freely falling particles decreases as $t \rightarrow 0$, the decrease is not rapid enough to insure that sufficiently distant regions of the universe were in causal contact. The FLRW models have particle horizons. Horizons in cosmology measure the maximum distance light travels within a given time period from a time of emission t_e ; the “particle horizon” is defined as the limiting case $t_e \rightarrow 0$.¹³ The existence of particle horizons in the FLRW models indicates that distant regions are not in causal contact (see Figure 9.1). Many discussions mistakenly refer to the Hubble radius H^{-1} as the “horizon.” This is a misnomer because the Hubble radius is not defined in terms of causal structure, but it does indicate the length scale at which expansion has an impact on evolution of perturbations. A simple scaling argument shows that in standard FLRW expansion, perturbation wavelengths cross the “horizon”: the perturbation wavelengths simply scale with the expansion, whereas H^{-1} scales as $H^{-1} \propto R^{1/n}$ for $R(t) \propto t^n$. For the length scale associated with a galaxy, horizon crossing occurs at around $t \approx 10^9$ seconds. It is puzzling that the perturbations are coherent prior to this time, at a length scale larger than the Hubble radius.

One response was to hope that new physics would lead to a different causal structure of the early universe. Bardeen concludes a study of the evolution of density perturbations as follows (Bardeen 1980, 1903):

The one real hope for a dynamical explanation of the origin of structure in the Universe is the abolition of particle horizons at early times, perhaps through quantum modifications to the energy-momentum tensor and/or the gravitational field equations which in effect violate the strong energy condition.¹⁴

¹³A horizon is the surface in a time slice t_0 separating particles moving along geodesics that could have been observed from a world line γ by t_0 from those which could not (Rindler 1956). The distance to this surface, for signals emitted at a time t_e , is given by

$$d = R(t_0) \int_{t_e}^{t_0} \frac{dt}{R(t)} \tag{9.1}$$

Different “horizons” correspond to different choices of limits of integration. The integral converges for $R(t) \propto t^n$ with $n < 1$, which holds for matter or radiation-dominated expansion. Thus the integral for the particle horizon ($\lim_{t_e \rightarrow 0}$) converges for the FLRW models (e.g., Ellis and Rothman 1993).

¹⁴Energy conditions are constraints on what is taken to be a reasonable source for the gravitational field equations. Roughly speaking, the strong energy condition requires that the stresses in matter will not be so large as to produce negative energy densities. Formally, $T_{ab}\xi^a\xi^b \geq \frac{1}{2}\text{Tr}(T_{ab})$ for every unit timelike ξ^a ; for a perfect fluid, this implies that $\rho + 3p \geq 0$, where ρ is the energy density and p is the pressure. As Bardeen notes, if the strong energy condition fails then there are solutions such that the integral in Equation (9.1) diverges.

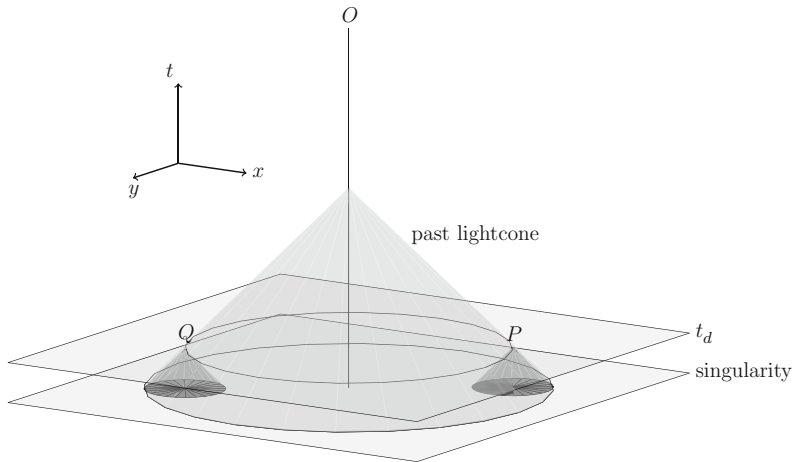


Fig. 9.1 This diagram, with light cones at 45° , illustrates the causal structure of the FLRW models. Points P , Q on the surface of last scattering t_d , both falling within the past light cone of an observer O , do not have overlapping light cones.

But Bardeen’s focus on particle horizons as a fundamental obstacle sets him apart from others in the field; Peebles (1980), for example, mentions the puzzles associated with horizons but apparently takes this to be one of the many indications that we do not sufficiently understand physics near the big bang.

The second puzzle regards the amplitude of the perturbations as they crossed the Hubble radius. While this could be treated as a parameter to be fixed by observations, many theorists hoped for a physical account of how this amplitude was fixed in the early universe. One can evolve backward to determine the amplitude of the fluctuation spectrum at a given “initial” time t_i . For t_i on the order of the Planck time, for example, these fluctuations are much *smaller* than thermal fluctuations, which are taken to be physically plausible.¹⁵ It seems inappropriate to treat t_i as a free variable, choosing when to “imprint” a spectrum of thermal fluctuations such that the amplitudes match observations. The Planck time is often singled out on dimensional grounds as the scale at which quantum gravity effects should become important. But in the absence of a successor theory, it is unclear how to delimit the boundary of applicability of classical GR and then choose a plausible “initial” perturbation spectrum.

By the late 1970s and early 1980s, several cosmologists had greater ambitions than merely giving a phenomenological account of structure formation. They sought to understand the origins of initial perturbations based on new physics applicable to the early universe. Those sharing this ambition could draw ideas from the ample

¹⁵For example, Blau and Guth (1987) compare the density contrast imposed at $t_i = 10^{-35}$ seconds to the fluctuations obtained by evolving backward from the time of recombination implies $\Delta \approx 10^{-49}$ at t_i , nine orders of magnitude *smaller* than thermal fluctuations.

storehouse of speculative physics: Planck scale metric fluctuations, gravitational particle production, primordial black holes, “gravithermal” effects, primordial turbulence, nonequilibrium dynamics, and so on.¹⁶ Sakharov (1966) was the first to propose a detailed quantum description of the initial perturbations—remarkably, before the discovery of the CMBR. But this early paper drew no attention, partially because it was an extension of Zel’dovich’s “cold bang” proposal that fell from favor following the discovery of the CMBR. From the mid-1970s onward, several theorists explored the implications of early universe phase transitions for structure formation, in particular the production of topological defects (discussed in more detail below). This work, along with studies of other possible impacts of phase transitions, illustrates that giving a physical account of the earliest stages of structure formation came to be regarded as a viable research topic. As of 1980 the field was wide open, with the potential to draw on ideas in general relativity and quantum gravity or the many novel ideas recently introduced in particle physics.

In addition to puzzles regarding the initial perturbations, these approaches to structure formation were threatened by tightening observational constraints based on the isotropy of the temperature of the CMBR. Partridge (1980) reached sensitivities of $\Delta T/T \approx 10^{-4}$ in isotropy measurements, and at this level he should have detected fluctuations according to either of the prevailing accounts of structure formation. This problem, along with other events such as experimental evidence in favor of a massive neutrino, led theorists to add hot and cold dark matter to their models of structure formation starting in the early 1980s (see, e.g., Peebles 1982 and Pagels 1984).¹⁷ The early dark matter models established the compatibility between the observational upper limits on temperature anisotropies in the CMBR and the idea of structure formation via gravitational instability. Adding cold dark matter helps to reconcile the uniformity of the CMBR with later clumpiness of matter because, roughly speaking, the cold dark matter decouples from the baryonic matter and radiation early, leaving a minimal imprint on the CMBR, yet after recombination the cold dark matter perturbations regenerate perturbations in the baryonic matter sufficiently large to seed structure formation.

Many contemporary textbooks on structure formation use the puzzles regarding initial perturbations described above to set the stage for the entrance of inflationary cosmology. Rather than pulling the initial spectrum out of a hat, as one might suspect of the earlier proposals, the inflationary theorist can pull an HPZ spectrum with an appropriate amplitude out of the vacuum fluctuations of a quantum field. The performance is captivating because it displays the possibility of *calculating* the features of the initial spectrum from physical principles. The following section will

¹⁶See Barrow (1980) for a brief review of some of these ideas and references and Peebles (1980); Zel’dovich and Novikov (1983) for more comprehensive overviews of the field.

¹⁷“Hot” vs. “cold” refers to the thermal velocities of relic particles for different types of dark matter. Hot dark matter decouples while still “relativistic,” in the sense that the momentum is much greater than the rest mass, and relics at late times would still have large quasi-thermal velocities. Cold dark matter is “nonrelativistic” when it decouples, meaning that the momentum is negligible compared to the rest mass and relics have effectively zero thermal velocities.

review the route by which the theorists discovered this appealing consequence of inflation and assess its importance by contrast with the other features of inflation emphasized by Guth (1981).

9.3 Inflationary Cosmology

The essential idea of inflation is that the early universe went through a transient phase of de Sitter-like expansion.¹⁸ During this phase the scale factor grows exponentially with time, $R(t) \propto e^{\chi t}$, compared to the more sedate radiation-dominated FLRW expansion with $R(t) \propto t^{1/2}$. The idea of modifying FLRW expansion in this way had been suggested several times prior to 1980 (see Smeenk 2005), and the earlier proposals shared two common problems. First, what is the physical source of the accelerated expansion? I will refer to this as the source problem. The source could not be garden-variety matter or radiation, because to drive a stage of exponential expansion it would have to violate the strong energy condition typically assumed to hold for reasonable matter fields.¹⁹ Second, how does the exponential expansion transition into the usual FLRW expansion with appropriate matter and energy densities? Solving this second problem, the transition problem, requires an explanation of how the physical source of the expansion ceased to be dynamically relevant and sets the stage for the standard big bang model. Any matter or radiation present at the onset of exponential expansion is rapidly diluted away, leaving only the vacuum energy ρ_v , which remains constant throughout the expansion. One needs an account of how the universe is re-populated with normal matter and radiation after the stage of exponential expansion.

Guth (1981) launched a research program not by solving either of these problems but by making a compelling case in favor of inflation. He recognized that a stage of exponential expansion solves two fine-tuning problems of the standard model, the flatness and horizon problems. On this basis he argued that the idea was worth pursuing despite his failure to give an account of the transition to the standard model. The source of exponential expansion in his original account was the vacuum energy of the Higgs field in a proposed grand unified theory (GUT) trapped in a

¹⁸De Sitter spacetime is a solution to Einstein's field equations with a stress-energy tensor given by $T_{ab} = -\rho_v g_{ab}$, where ρ_v is the vacuum energy density. The scale factor then expands exponentially, with $\chi^2 = \frac{8\pi}{3}\rho_v$. During inflation the stress-energy tensor has approximately this form. Given that the vacuum energy density remains constant during the expansion while "ordinary" matter and energy is rapidly diluted, the vacuum energy dominates the expansion and the solution, roughly speaking, approaches de Sitter spacetime.

¹⁹The stress-energy tensor stated in the previous footnote does not satisfy the strong energy condition formulated in footnote 14; the fact that the vacuum energy density does not dilute with expansion reflects this. A stress-energy tensor that violates this condition is a necessary condition for exponential expansion within classical GR.

false minima during a first-order phase transition.²⁰ Even though this solution of the source problem would not survive long, by contrast with earlier proposals, Guth had shown how to link the idea of inflation with an active area of research in particle physics. In effect, inflation exchanged various large-scale properties of the universe, previously treated as initial conditions, for features of the dynamical evolution of a scalar field in the early universe. This exchange was soon exploited in giving a solution to the transition problem and in giving an account of the origins of the seeds for structure formation. After reviewing Guth's case and critical responses to it, we will turn to the discovery of the inflationary account of structure formation at the Nuffield workshop and briefly discuss the account itself in more detail.

9.3.1 *Inflation as a Theory of Initial Conditions*

Guth identified two problems that inflation was able to solve:

The standard model of hot big-bang cosmology requires initial conditions which are problematic in two ways: (1) The early universe is assumed to be highly homogeneous, in spite of the fact that separated regions were causally disconnected (horizon problem) and (2) the initial value of the Hubble constant must be fine tuned to extraordinary accuracy ... (flatness problem). (Guth 1981, 347)

The first could be more aptly called the “uniformity problem”: there is an apparent conflict between the strikingly uniform temperature of the CMBR and the horizon structure of the FLRW models (see Figure 9.1). We have seen above that cosmologists working on structure formation noted puzzles due to horizons, and Misner (1969) formulated the problem in terms similar to Guth's a decade earlier.²¹ The portion of the universe we can see consists of 10^{83} causally disconnected regions at the Planck time. Due to horizons, the fact that all these regions have the same physical properties cannot be explained via causal interactions.

What Guth called the “flatness problem” had not been widely discussed.²² The dynamics of the FLRW models implies that all models approach the “flat” model at early times. This can be seen in the behavior of the density parameter Ω , which is

²⁰Guth discovered inflation while focusing on a third problem, the monopole problem. GUTs from the late 1970s predicted the existence of magnetic monopoles, and the relic abundance of the monopoles would be many orders of magnitude greater than the observed energy density of the universe. See Guth (1997a) for his account of how he discovered inflation. Unlike the monopole problem, which arises for the combination of cosmology and these GUTs, the flatness and horizon problems are problems for the cosmological standard model.

²¹See Smeenk (2005) for a discussion of how these two features of the FLRW models were treated prior to Guth's identification of them as problems to be solved by inflation.

²²Guth learned of the problem from lectures given by Robert Dicke (Guth 1997a). See Dicke (1969) and Dicke and Peebles (1979) for Dicke's formulation of the problem, which he characterized as an “enigma.”

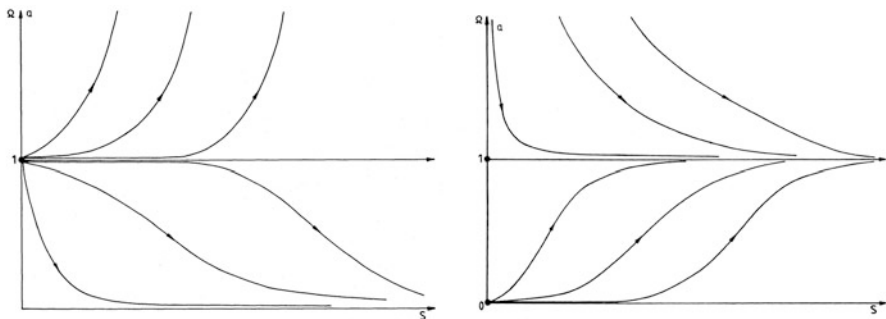


Fig. 9.2 These diagrams from Ellis and Madsen (1988) illustrate the evolution of Ω as a function of S (the scale factor R in the text). In the left diagram, Ω diverges from 1 (for $\gamma > 2/3$), whereas on the right Ω is driven toward one during an inflationary phase (with $\gamma = 0$).

1 for the FLRW model with Euclidean spatial sections.²³ Assuming normal matter and radiation as sources, $|\Omega - 1|$ increases with time under the FLRW dynamics.²⁴ It is thus surprising to discover that $\Omega(t_0)$, the current observed value, is quite close to 1. If we imagine choosing a value $\Omega(t_i)$ at some early time, it must be *incredibly* close to 1 to be compatible with observations (see Figure 9.2).

Thus the standard model requires positing, at the Planck time, the same physical conditions in 10^{83} causally disconnected patches, with a delicately chosen total energy density. Guth argued that inflation is compatible with a much more plausible initial state. Inflation stretches the horizon length; for N “e-foldings” of expansion, the horizon length d_h is multiplied by e^N . For $N > 65$ the horizon distance, while still finite, encompasses the observed universe. The observed universe could then have evolved from a single causal patch rather than 10^{83} patches with an astonishing degree of pre-established harmony. In addition, during inflation the density parameter is driven *toward* one.²⁵ An inflationary stage long enough to solve the horizon problem drives a large range of pre-inflationary values of $\Omega(t_i)$ sufficiently close to 1 by the end of inflation, such that $\Omega(t_0) \approx 1$. If inflation occurs,

²³ $\Omega =: \frac{\rho}{\rho_c}$, where the critical density ρ_c is the value required for the attraction of gravity due to positive matter-energy density to precisely balance the initial expansion and cosmological constant: $\rho_c = \frac{3}{8\pi} (H^2 - \frac{\Lambda}{3})$.

²⁴More precisely,

$$\frac{|\Omega - 1|}{\Omega} \propto R(t)^{3\gamma-2}, \tag{9.2}$$

where γ is used to classify different types of perfect fluids. The equation of state of a perfect fluid is $p = (\gamma - 1)\rho$, where p is the pressure and ρ the density. For radiation, $\gamma = 4/3$ and for “dust” $\gamma = 1$ (corresponding to zero pressure). For “normal” matter, satisfying the energy conditions defined in footnote 14, $\gamma > 2/3$.

²⁵During inflation, the strong energy condition is violated, and $\gamma = 0$; it is clear from Equation (9.2) that Ω is then driven toward 1.

there is no need for a finely tuned choice of $\Omega(t_i)$. A word of caution is in order: however, it is not the case that inflation *eliminates* dependence on initial conditions entirely. One can choose initial conditions that lead to an arbitrarily nonuniform universe with any value of Ω , despite inflation's preference for a uniform universe with $\Omega(t_0) = 1$. Inflation *enlarges* the range of initial conditions compatible with observations.²⁶

For those working, like Guth, in particle physics, these problems and the approach to solving them had a familiar ring. Following Wilson and 't Hooft, many particle theorists sought extensions of the standard model of particle physics that would eliminate its “unnatural” features, such as the huge discrepancy between the Higgs boson mass and the scale of the fundamental interactions (the Planck scale). The reception of Guth's case for inflation in the particle physics community reflects, in part, acceptance of a common strategy: using fine-tuning as a guide to developing new theories. The reception among astrophysicists and cosmologists was more uneven. These communities did not share the methodology implicit in focusing on naturalness or fine-tuning problems. Brawer (1996) argues that Guth's discovery of a solution to the horizon and flatness problems helped to convince many that they were, in fact, legitimate problems.²⁷ But there have been, since inflation was introduced, vocal critics who have rejected the idea that inflation should be given credence based on solving fine-tuning problems.

One line of criticism grants that an early universe theory should explain how the observed universe arose from “generic” initial conditions, as Guth argued. But does inflation deliver such an explanation? Penrose (1986) argued that the probability of inflation must itself be quite low, on general grounds.²⁸ Suppose that we are given a generic state in a universe that evolves into a “big crunch” singularity in the future. It seems overwhelmingly unlikely that as the universe approaches the final singularity, it will “deflate” by converting all the gravitational energy of the collapsing matter into kinetic energy of a scalar field in just the right way to push

²⁶This point was first made in response to Misner's “chaotic cosmology,” which like inflation proposed new dynamics (in Misner's case, damping of anisotropies due to neutrino viscosity) in order to insure that an isotropic universe emerges from a large range of anisotropic initial conditions. In response to Misner, Collins and Stewart (1971) showed that one can always pick an arbitrarily large anisotropy at a given time t_0 and find a solution of the relevant system of equations as long as there are no processes which could prevent arbitrarily large anisotropies at some $t_i < t_0$. A similar criticism applies to inflation, as Madsen and Ellis (1988) have emphasized. Guth (1997b) has acknowledged this point: “. . . I emphasize that *NO* theory of evolution is ever intended to work for arbitrary initial conditions. . . . In all cases, the most we can hope for is a theory of how the present situation could have evolved from *reasonable* initial conditions” (pp. 240–241, emphasis in the original).

²⁷Brawer's case is based on published discussions of these problems, as well as the extensive interviews with cosmologists published in Lightman and Brawer (1990).

²⁸Penrose's original terse statement of this criticism appeared in a book review of the conference proceedings of the Nuffield workshop (discussed below), and he has discussed it further in Penrose (1989, 2004). Although I do not have the space to discuss Penrose's objection, several more recent papers pursue the issues raised by Penrose, including Unruh (1997) and Hollands and Wald (2002).

it into a false vacuum state. But this is simply the time reverse of the account of inflation, and—on the assumption that the dynamics is time reversal invariant—the argument concludes that our assessment that the probability of deflation is low should also apply to inflation itself. If Penrose is correct, the features of the field driving inflation, and its pre-inflationary state, are more finely tuned than the initial state required in the standard model without inflation. Inflation shifts fine-tuning from one place to another rather than eliminating it. (There are further specific constraints required on the pre-inflationary state. Vachaspati and Trodden (1999) proved that the field driving inflation must be uniform over a region larger than the Hubble radius in order to trigger inflation; and the scalar field has to be sufficiently uniform to drive exponential expansion during inflation.²⁹)

A second line of argument questions why we should grant that the universe began in a “generic” initial state.³⁰ This is, in effect, a hypothesis regarding the universe at the Planck scale, which is uncertain due to our lack of a theory of quantum gravity. What grounds do we have for accepting such a hypothesis? One problem regards even formulating the hypothesis. Guth and others often write as if the initial state should be regarded as “chosen at random” from among a set of possibilities. It is unclear, however, what theory should be used to define the space of possibilities, since classical GR does not adequately reflect all the laws that would govern this domain. Furthermore, the assessment of an initial state as “generic,” or, on the other hand, “special,” is based on a choice of measure over the allowed initial states of the system. But on what grounds is one measure to be chosen over another? Even if we obtain a clearly delimited space of possibilities, equipped with a measure that allows us to determine the properties of a “generic” choice, what justifies this hypothesis? For a normal experimental system, it is possible to check, at least in principle, whether a large variety of initial states lead to the same final state; if so, there is evidence that the system is governed by dynamics that washes away dependence on the initial state. Obviously, however, such supporting evidence cannot be gathered in cosmology.

A quite different response is that Guth’s motivations for inflation should be disentangled from the physics. In fact, Guth’s precursors in the Soviet Union introduced inflation with essentially the opposite methodology. For example, Starobinsky

²⁹The stress-energy tensor for a scalar field is given by

$$T_{ab} = \nabla_a \phi \nabla_b \phi - \frac{1}{2} g_{ab} \left(g^{cd} \nabla_c \nabla_d \phi - V(\phi) \right); \quad (9.3)$$

inflation requires that the field is “potential-dominated” in the sense that the field is sufficiently uniform that the derivative terms are negligible, $V(\phi) \gg g^{cd} \nabla_c \nabla_d \phi$. If this condition holds, $T_{ab} \approx -V(\phi) g_{ab}$ as required to produce exponential expansion.

³⁰Different versions of this line of argument have been pressed by a number of critics of inflation; see, for example, Earman and Mosterin (1999), Hollands and Wald (2002), and more recently Ijjas et al. (2013).

(1978, 1979) regarded the choice of a specific initial state—de Sitter spacetime—as extremely natural, and it had the advantage of evading the singularity in FLRW models.

Critical responses along these lines did not dim the enthusiasm for Guth’s proposal. This is, I will argue, in part because of the answer that was shortly developed to a problem that Guth did see as a clear obstacle to the idea of inflation. Guth noted the advantages of inflation while at the same time admitting that his model failed to solve the transition problem (also called the graceful exit problem). Rather than smoothly joining onto the FLRW expansion, the phase transition Guth considered ended via bubble nucleation, leaving the early universe marred with nonuniformities. The model failed to achieve the delicate balance between overall uniformity and slight perturbations required for the account of structure formation via gravitational instability. As Barrow and Turner (1981) noted, at first blush, and provided that bubble nucleation could be avoided, inflation may actually exacerbate the problem by too efficiently smoothing out the universe, leaving it without wrinkles to seed later structures. This worry led to an important success as theorists discovered a mechanism for generating perturbations during inflation.

9.3.2 *New Inflation and the Nuffield Workshop*

Guth’s paper and talks based on it introduced many astrophysicists and particle physicists to the very idea of early universe cosmology. By admitting the flaws of his initial model, Guth also left his readers and audiences with a project: to find a working model of inflation. Paul Steinhardt, then a junior fellow in the Harvard Society of Fellows, exemplifies this reaction; he described Guth’s talk at Harvard as “the most exciting and depressing talk” he had ever attended (Steinhardt 2002). The excitement stemmed from the promise of connecting the study of phase transitions to fundamental questions in cosmology. But after laying out inflation’s ability to solve the flatness, horizon, and monopole problems, Guth ended by explaining the fatal flaw of his initial model. Steinhardt recalls his reaction (Steinhardt 2002): “Here was this great idea and it just died right there on the table. So I couldn’t let that happen.”

Given Steinhardt’s background in condensed matter physics and familiarity with phase transitions, he was ideally suited to take on the task of reviving Guth’s idea. News of Guth’s paper also led Andrei Linde in Moscow, a pioneer in the study of early universe phase transitions throughout the 1970s, to reconsider the possibility of a first-order phase transition. Linde had considered the idea in collaboration with Chibisov but had dismissed it as unworthy of publication—“there was no reason to publish such garbage”—due to the problem of inhomogeneities.³¹ Steinhardt began

³¹The collaborative work with Chibisov is mentioned in Linde (1979, 433–434); the quotation is from a 1987 interview (Lightman and Brawer 1990, 486–486).

studying early universe phase transitions almost immediately, and upon taking a faculty position at the University of Pennsylvania, he found a graduate student, Andy Albrecht, eager to join in the project. Linde and Steinhardt and Albrecht independently realized that a symmetry-breaking phase transition governed by a different effective potential than that used by Guth could solve the transition problem while providing sufficient inflation to solve the horizon and flatness problems (Albrecht and Steinhardt 1982; Linde 1982). Their proposal is usually called “new inflation.”³²

Albrecht and Steinhardt (1982) and Linde (1982) both developed models of the phase transition based on a Coleman-Weinberg effective potential for the Higgs field. (The Lagrangian density for a classical scalar field is given by $\mathcal{L} = \frac{1}{2}\partial_\mu\phi\partial^\mu\phi - V(\phi)$, where $V(\phi)$ is the potential. The effective potential includes quantum corrections to the classical potential.)³³ This change leads to a dramatically different phase transition. Most importantly, inflation continues after the formation of an initial bubble: rather than tunneling directly to the global minimum, in this scenario the field ϕ evolves to the minimum over a “long” time scale τ (i.e., much longer than the expansion time scale). Throughout this evolution, ϕ is still displaced from the global minimum, and the nonzero $V(\phi)$ continues to drive exponential expansion. Linde (1982) and Albrecht and Steinhardt (1982) both argue that for natural values of τ , the expansion lasts long enough for the initial bubble to become much larger than the observed universe. Finally, as in Guth’s scenario any pre-inflationary matter and energy density are diluted during the extended inflationary stage. In the new scenario, oscillations of the field ϕ near its global minimum would produce other particles via baryon-number-nonconserving decay in order to “reheat” the universe to an energy density compatible with standard cosmology.

The initial proposals were quickly developed into a general account of new inflation. The features of the phase transition can be described simply in terms of the evolution of ϕ , which is determined by the form of the potential $V(\phi)$. The classical equations of motion for a scalar field ϕ with a potential $V(\phi)$ in an FLRW model are given by

³²At roughly the same time, Stephen Hawking and Ian Moss proposed an alternative solution to the transition problem. Although Hawking and Moss (1982) is sometimes cited as a third independent discovery of new inflation, it differs substantially from the other proposals. The aim of the paper is to show that including the effects of curvature and finite horizon size leads to a different description of the phase transition. This phase transition proceeds from a local minimum at $\phi = 0$ to the global minimum ϕ_0 via an intermediate state ϕ_1 ; rather cryptic arguments lead to the conclusion that “the universe will continue in the essentially stationary de Sitter state until it makes a quantum transition everywhere to the $\phi = \phi_1$ solution” (p. 36). They further argue that following this transition to a coherent Hubble scale patch, ϕ will “roll down the hill” (for an appropriate values of parameters in the effective potential), producing an inflationary stage long enough to match Guth’s success.

³³See, e.g., Coleman (1985), Chapter 5 for an introduction to the effective potential, and Kolb and Turner (1990) for a detailed discussion of the differences between old and new inflation.

$$\frac{d^2\phi}{dt^2} + 3H\frac{d\phi}{dt} + \Gamma_\phi\frac{d\phi}{dt} + \frac{dV(\phi)}{d\phi} = 0, \quad (9.4)$$

where t is the time coordinate in the FLRW model and Γ_ϕ is the decay width of ϕ .³⁴ New inflation requires a long “slow roll” followed by reheating. Assume that the field ϕ is initially close to $\phi = 0$. Slow roll occurs if the potential is suitably flat near $\phi = 0$ and the $\ddot{\phi}$ term is negligible; given the further assumption that the Γ_ϕ term is negligible, then the evolution of ϕ can be approximately described by

$$3H\dot{\phi} \approx -\frac{dV(\phi)}{d\phi}. \quad (9.5)$$

(The name is due to the similarity between the evolution of ϕ and that of a ball rolling down a hill, slowed by friction.) During slow roll, the potential energy $V(\phi)$ dominates over the kinetic energy $\frac{\dot{\phi}^2}{2}$, and $V(\phi)$ drives inflationary expansion. The slow roll approximation breaks down as the field approaches the global minimum. The Γ_ϕ term is put in “by hand” to describe the process of reheating: roughly, ϕ oscillates around the minimum and decays into other types of particles. The details depend on the coupling of ϕ to other fields and are heavily model-dependent. The reheating stage is necessary to “repopulate” the universe, given that any preexisting matter or radiation is rapidly diluted during the inflationary expansion.

By the spring of 1982, several groups were at work fleshing out the details of the new inflationary scenario: a group at the University of Chicago and Fermilab including Turner and Kolb, Steinhardt and Albrecht at the University of Pennsylvania, Guth at MIT, Linde and various collaborators in Moscow, Laurence Abbott at Brandeis, Hawking and others in Cambridge, and John Barrow in Sussex. With notable exceptions, such as Hawking and Barrow, nearly everyone in this research community came from a background in particle physics. The framework described in the previous paragraph left ample room for innovation and new ideas: the connections with particle physics were poorly understood at best, the various approximations used were generally on shaky footing, and there were numerous hints of interesting new physics. Several of these researchers recognized the most important hint: homogeneity at all scales at the end of inflation would be incompatible with accounts of galaxy formation, which required an initial spectrum of perturbations. There appeared to be several ways to avoid *too much* homogeneity at the end of inflation; Linde (1982), for example, mentions a later phase transition without supercooling or quantum gravity effects as a possible means for generating inhomogeneities.

³⁴One of the main differences between the initial papers on new inflation is that Albrecht and Steinhardt (1982) explicitly include the $3H\dot{\phi}$ term (aka the “Hubble drag” term), whereas Linde (1982) does not.

The first international conference focusing on “very early universe cosmology ($t < 1$ sec)” convened in Cambridge from June 21–July 9, 1982.³⁵ Nearly half the lectures at the Nuffield workshop were devoted to inflation, and the intense collaborations and discussions during the workshop led to the “death and transfiguration” of inflation (from the title of the conference review in *Nature* (Barrow and Turner 1982)). One focus of the conference was the calculation of density perturbations produced during an inflationary stage: Steinhardt, Starobinsky, Hawking, Turner, Lukash, and Guth had all realized that this was a “calculable problem” (in Steinhardt’s words), with the answer being an estimate of the magnitude of the density perturbations, measured by the dimensionless density contrast Δ , produced during inflation. Preliminary calculations of this magnitude disagreed by an astounding 12 orders of magnitude: Hawking circulated a preprint (later published as Hawking 1982) that found $\Delta \approx 10^{-4}$, whereas Steinhardt and Turner (1984) initially estimated a magnitude of 10^{-16} . After 3 weeks of effort, the various groups working on the problem had converged on an answer, but the answer proved to be disastrous for new inflation.

The calculations drew on an idea introduced prior to Guth’s paper. Mukhanov and Chibisov (1981) had argued that a de Sitter phase could generate perturbations by “stretching” zero-point fluctuations of quantum fields to significant scales. This idea would become the basis for the generation of seed perturbations in inflationary cosmology. The details were worked out at the Nuffield workshop, which seems to be a rare example of a scientific workshop that fulfilled the goal of bringing together the relevant research groups and successfully forging a consensus on an important problem.

Prior to the workshop, Hawking circulated a preprint which argued that initial inhomogeneities in the ϕ field would imply that inflation begins at slightly different times in different regions; the inhomogeneities reflect the different “departure times” of the scalar field. Hawking’s preprint claimed that this results in a scale-invariant spectrum of adiabatic perturbations with $\Delta \approx 10^{-4}$, exactly what was needed in accounts of structure formation. But others pursuing the problem (Steinhardt and Turner; Guth and his collaborator, So-Young Pi) did not trust Hawking’s method; Steinhardt has commented that he “did not believe it [Hawking’s calculation] for a second” (Steinhardt 2002, cf. Guth 1997a, 222–230). There were two closely-linked concerns with Hawking’s method (beyond the sketchiness of his initial calculations): it is not clear how this approach treats the evolution of the fluctuations in different regimes, and it is also not gauge invariant.

The “gauge problem” in this case reflects the fact that a “perturbed spacetime” cannot be uniquely decomposed into a background spacetime plus perturbations.

³⁵The description is taken from the invitation letter to the conference (Guth 1997a, 223). The Nuffield Foundation had previously sponsored conferences in quantum gravity but shifted the focus to early universe cosmology in response to interest in the inflationary scenario. A 1981 conference in Moscow on quantum gravity also included numerous discussions of early universe cosmology (Markov and West 1984), but Nuffield was the first conference explicitly devoted to the early universe.

Slicing the spacetime up along different surfaces of constant time leads to different magnitudes for the density perturbations. The perturbations “disappear,” for example, by slicing along surfaces of constant density. In practice, almost all studies of structure formation used a particular gauge choice (synchronous gauge), but this leads to difficulties in interpreting perturbations with length scales greater than the Hubble radius.³⁶ Press and Vishniac (1980) identify six “tenacious myths” that result from the confusion between spurious gauge modes and physical perturbations for $\lambda > H^{-1}$. This problem is significant for the inflationary account because over the course of an inflationary stage perturbations of fixed length go from $\lambda \ll H^{-1}$ to $\lambda \gg H^{-1}$. Length scales “blow up” during inflation since they scale as $R(t) \propto e^{Ht}$, but the Hubble radius remains fixed since H is approximately constant during the slow-roll phase of inflation. For this reason, it is especially tricky to calculate the evolution of physical perturbations in inflation using a gauge-dependent formalism. The first problem mentioned in the previous paragraph is related: determining the imprint of initial inhomogeneities requires evolving through several regimes, from the pre-inflationary patch through the inflationary stage and reheating to standard radiation-dominated evolution.

Hawking and Guth pursued refinements of Hawking’s approach throughout the Nuffield workshop.³⁷ The centerpiece of these calculations is the “time delay” function characterizing the start of the scalar field’s slow roll down the effective potential. This “time delay” function is related to the two-point correlation function characterizing fluctuations in ϕ prior to inflation, and it is also related to the spectrum of density perturbations, since these are assumed to arise as a result of the differences in the time at which inflation ends. However, these calculations treat the perturbations as departures from a globally homogenous solution to the equations of motion for ϕ and do not take gravitational effects into account. How this approach is meant to handle the gauge problem is also not clear.

Starobinsky’s approach leads to a similar conclusion via a different argument: as in the first approach, the time at which the de Sitter stage ends is effectively coordinate dependent (Starobinsky 1982). The source of these differences is traced to the production of “scalars” during the de Sitter stage rather than a “time delay” function for the scalar field (see, in particular, Starobinsky 1983, 303). Finally, Steinhardt and Turner enlisted James Bardeen’s assistance in developing a third approach; he had recently formulated a fully gauge invariant formulation for the study of density perturbations (Bardeen 1980). Using Bardeen’s formalism, the three aimed to give a full account of the behavior of different modes of the field ϕ as these evolved through the inflationary phase and up to recombination. The physical origin of the spectrum was traced to the qualitative change in behavior as perturbation

³⁶Synchronous gauge is also known as “time-orthogonal” gauge: the coordinates are adapted to constant time hypersurfaces orthogonal to the geodesics of comoving observers. All perturbations are confined to spatial components of the metric, i.e., the metric has the form $ds^2 = R^2(t)(dt^2 - h_{ij}dx^i dx^j)$, with $i, j = 1, 2, 3$. The coordinates break down if the geodesics of co-moving observers cross.

³⁷These efforts were later published as Hawking (1982) and Guth and Pi (1982).

modes expand past the Hubble radius: they “freeze out” as they cross the horizon and leave an imprint that depends on the details of the model under consideration.

Here I will not give a more detailed comparison of these three approaches. Despite the conflicting assumptions and other differences, the participants of the Nuffield workshop apparently lent greater credibility to their conclusions due to the rough agreement between the three different approaches.

During the 3 weeks of collaboration at Nuffield, these different approaches converged on the following results. In the notation of Bardeen et al. (1983), the spectrum of density perturbations is related to the field ϕ by

$$\Delta|_{\lambda} = AH \frac{\Delta\phi}{\dot{\phi}}, \quad (9.6)$$

where $\lambda \approx H^{-1}$ and A is a constant depending on whether the universe is radiation ($A = 4$) or matter ($A = 2/5$) dominated when λ “reenters” the Hubble radius. The other quantities on the RHS are both evaluated when λ “exits” the Hubble radius: $\Delta\phi$ is the initial quantum fluctuation in ϕ , on the order of $\frac{H}{2\pi}$. The value of $\dot{\phi}$ is given by (from (9.5)) $\dot{\phi} \approx \frac{V'(\phi)}{3H}$, and V' depends on the coupling constants appearing in the effective potential. For a Coleman-Weinberg effective potential with “natural” coupling constants, $\dot{\phi} < H^2$; plugging this all back into the initial equation we have

$$\Delta|_{\lambda} > A \frac{H^2}{2\pi H^2} \approx .1 - 1 \quad (9.7)$$

Inflation naturally leads to an *almost* HPZ spectrum, which is also Gaussian (see, e.g., Bardeen et al. 1983). But reducing the magnitude of these perturbations to satisfy observational constraints requires an unnatural choice of coupling constants. In particular, the self-coupling for the Higgs field apparently needs to be on the order of 10^{-8} , although a “natural” value would be on the order of 1.³⁸

Calculations of the perturbation spectrum culminated in a Pyrrhic victory: a Coleman-Weinberg potential provided a natural mechanism for producing perturbations, but it could be corrected to give the correct amplitude only by abandoning any pretense that the field driving inflation is a Higgs field in an $SU(5)$ GUT. However, it was clear how to develop a “newer inflation” model; even before the conclusion of the conference, Bardeen, Steinhardt, and Turner had suggested that the effective potential for a scalar field in a supersymmetric theory (rather than the Higgs field of a GUT) would have the appropriate properties to drive inflation.

Finding a particular particle physics candidate for the scalar field driving inflation would provide for an important independent line of evidence. The Nuffield workshop marked the start of a different approach, as the focus shifted to implementing

³⁸See Steinhardt and Turner (1984, 2165–2166) for a clear discussion of this constraint, which is also discussed in detail in Kolb and Turner (1990); Linde (1990).

inflation successfully rather than starting with a candidate for the field driving inflation derived from particle physics.

The introduction of an “inflaton” field, a scalar field custom-made to produce an inflationary stage, roughly a year later illustrates this methodological shift.³⁹ An inflaton field may resemble the Higgs, but the rules of the game have changed: an inflaton is a new fundamental field distinct from any scalar field appearing in particle physics. The fact that inflation has not been closely tied to $SU(5)$ GUTs has been a boon to the field. Experiments carried out throughout the early to mid-1980s failed to detect proton decay on time scales predicted by the minimal $SU(5)$ GUTs (Blewitt et al. 1985). Following the demise of the minimal GUTs, there has been an ongoing effort to implement inflation within new models provided by particle physics.

Following the Nuffield workshop, inflation turned into a “paradigm without a theory,” borrowing Turner’s phrase, as cosmologists developed a wide variety of models bearing a loose family resemblance. The models share the basic idea that the early universe passed through an inflationary phase but differ on the nature of the “inflaton” field (or fields) and the form of the effective potential $V(\phi)$. Keith Olive’s review of the first decade of inflation ended by bemoaning the ongoing failure of any of these models to renew the strong connection with particle physics achieved in old and new inflation:

A glaring problem, in my opinion, is our lack of being able to fully integrate inflation into a unification scheme or any scheme having to do with our fundamental understanding of particle physics and gravity. . . . An inflaton as an inflaton and nothing else can only be viewed as a toy, not a theory. (Olive 1990, 389)

In a similar vein, Dennis Sciama commented that inflation had entered “a Baroque state” as theorists constructed increasingly ornate toy models (Lightman and Brawer 1990, p. 148). The sheer number of versions of inflation is incredible; Guth (1997a, 278) counts over 50 models of inflation in the nearly 3,000 papers devoted to inflation (from 1981 to 1997), and both numbers have continued to grow. Cosmologists have even complained about the difficulty of christening a new model with an original name, and a partial list of the inflationary menagerie has been used as comic relief in conference talks.⁴⁰

³⁹Several researchers studied scalar fields with the appropriate properties to drive inflation, but the term “inflaton” seems to have appeared first in Nanopoulos et al. (1983); see Shafi and Vilenkin (1984) for a similar model. I thank Keith Olive for bringing the first paper to my attention.

⁴⁰Rocky Kolb used such a slide in a talk at the Pritzker Symposium (Chicago, 1998); for an example of such a list, see Shellard (2003, figure 41.3).

9.4 Demise of a Rival Approach: Topological Defects

The development of scientific theories is shaped by competing approaches and the prospects for fruitful engagement with observations or experiments. The CMBR, aptly called the “cosmic Rosetta stone,” has provided a stable target for early universe cosmologists: the physical understanding of the CMBR is well-established, and the quality and variety of observations have improved steadily. These observations have been guided by assessments of what distinguishes among inflation and alternative accounts of the early universe. Throughout the 1980s and 1990s, the most important alternative account of the origins of structure was based on topological defects. These ideas were first studied in the 1970s prior to inflation, as a general feature of symmetry-breaking phase transitions in the early universe. Guth invented inflation to avoid an overabundance of one kind of defect, monopoles. But there are many types of defects that can be produced, and several theorists took up the challenge of understanding whether defects formed in the early universe could produce appropriate seeds for structure formation. This line of work is too diverse to be characterized as a single competing theory; it is, instead, a general approach, characterized by the assumption that topological defects are the primary mechanism of structure formation in the early universe. This brief discussion will focus on contrasting inflation with this approach, with no attempt to give a detailed account of the historical development of these ideas.⁴¹

The formation of topological defects is determined by properties of the vacuum manifold \mathcal{M} . The vacuum manifold consists of the degenerate vacuum states of the system after the phase transition. Suppose the theory initially has a symmetry group G that is then spontaneously broken to a subgroup H .⁴² The symmetry is broken in the sense that the vacuum states of the theory are degenerate: although the vacuum state is not invariant under the action of some $g \in G$, these distinct vacuum states are degenerate in that the Hamiltonian has the same eigenvalue. The subgroup H consists of those elements of G under which the vacuum state remains invariant. The space of degenerate vacuum states is then in one-to-one correspondence with sets of elements of the form gH ; in other words, the vacuum manifold \mathcal{M} is topologically

⁴¹I have left aside one important aspect of the comparison between inflation and topological defect theories, namely, the role of different types of dark matter in each scenario. The mechanisms for structure formation are part of package deal, including assumptions about the overall matter budget and other factors more significant for later stages of structure formation.

⁴²This means that, roughly speaking, for all $g \in G$, the Hamiltonian of the system is invariant under the action of g , but the vacuum or ground state of the system is not. (This is only a rough gloss; in quantum mechanics the action of a symmetry g is usually represented by a unitary operator on the Hilbert space, but in the case of broken symmetry, there is not a well-defined operator mapping between degenerate vacua, as these each define different Hilbert spaces.) The degenerate vacuum states are labeled by different values of the “order parameter” of the transition. The order parameter is the thermodynamic quantity that changes discontinuously through the transition and characterizes different phases, corresponding to degenerate vacua in this case; it is the vacuum expectation value of the relevant field(s).

equivalent to the quotient space G/H . Topological features of the vacuum manifold then determine what kinds of topological defects may form in the course of the phase transition.⁴³

Starting in the early 1970s, these ideas were applied to cosmology. Extrapolating the FLRW models, the early universe reaches arbitrarily high temperatures at early times. Kirzhnits (1972) suggested that symmetries in particle physics would be restored at sufficiently high temperatures, by analogy with symmetry restoration in condensed matter systems. Further calculations of symmetry restoration in the standard model of particle physics supported the idea that as the universe cooled, it passed through a series of phase transitions that broke the symmetries between various interactions. Many symmetry-breaking phase transitions in condensed matter systems lead to the formation of topological defects, such as vortices in liquid helium, so it is natural to expect that defects also arise in early universe phase transitions.

In a seminal paper, Kibble (1976) argued that topological defects would be produced due to the horizon structure of the early universe. (His account is sometimes referred to as the “Kibble mechanism.”) Given that the correlation length of the order parameter is bounded by the horizon distance, the phase transition produces domains in which the order parameter takes on different values determined by random fluctuations, assuming that the dynamics is not completely adiabatic. Whether defects form depends on the topology of the vacuum manifold. For example, suppose that there is a curve through \mathcal{M} that cannot be smoothly contracted to a point. Each point within the space \mathcal{M} represents a different degenerate vacuum state, which is labeled by different values of the order parameter for the phase transition.

Suppose that the values of the order parameter around a spatial loop take the same values given along the loop in \mathcal{M} . Since the loop cannot be continuously contracted to a point within \mathcal{M} , it is also not possible to assign values of the order parameter continuously in the region bounded by the spatial loop while remaining in \mathcal{M} . This implies that there must be a “defect,” namely, a region of space in which the fields cannot reach the vacuum state and instead remain trapped in a state of higher energy. The nature of these regions of higher energy is fixed by the structure of \mathcal{M} . In the case at hand, with a non-simply connected vacuum manifold, the phase transition leads to two-dimensional defects called “cosmic strings.” There are several other possibilities. A phase transition breaking a *discrete* symmetry leads to regions in which the order parameter takes on discrete values separated by domain walls, which are three-dimensional surfaces in spacetime. If the vacuum manifold has non-contractible two spheres rather than circles, then the phase transition produces

⁴³The relevant structure is given by the homotopy groups of the space. For further discussion, see, e.g., Vilenkin and Shellard (2000).

point-like defects (such as magnetic monopoles); for non-contractible three spheres, the corresponding zero-dimensional defects are called “textures,” event-like defects that do not have a stable localized core.⁴⁴

Early studies showed that domain walls and some types of monopoles had disastrous consequences, conflicting with observational constraints by several orders of magnitude (see, e.g., Zel’dovich et al. (1975), Zel’dovich and Khlopov (1978) and Guth and Tye (1980)). However, other types of defects—in particular, cosmic strings—were more plausible candidates for the seeds for structure formation. The defects are inherently stable regions of higher energy density, whose scale is set by the energy scale of the phase transition. The defects have an important impact on the dynamical evolution of the system following the phase transition, and in particular it is plausible that they will provide seeds that are subsequently enhanced via gravitational instability as described by linear perturbation theory. For GUT-scale phase transitions, the energy density is the appropriate order of magnitude to seed large-scale structure. Some defect theories have “scaling solutions,” in which the network of defects evolves such that there is no preferred length scale imprinted at a particular time. These theories then pass an important initial test, in that they lead to an approximately scale-invariant HPZ spectrum of perturbations.⁴⁵ They are thus compatible with the first generation of CMBR observations and the general picture of structure formation described above. However, there are important general differences between the inflationary account and that provided by topological defects, and these were clarified by a substantial research effort throughout the 1980s and 1990s.

To determine whether topological defects suffice as the primary mechanism for producing seeds for structure formation, researchers had to tackle two challenging problems. The first was to describe the phase transition itself and determine the nature of the defects produced, with sufficient quantitative detail to determine the consequences for later stages of evolution. In principle these details should be calculable given a particular extension of the standard model of particle physics. But the sheer complexity of the models, and the nature of the quantities needed to assess the implications for structure formation, has made it quite difficult in practice to carry out such calculations. Second, one had to describe the subsequent evolution of the network of defects leftover following the phase transition over a wide range of dynamical scales. Solving this second problem requires determining

⁴⁴Additional types of defects arise due to the distinction between gauge and global symmetries and the possibility of “hybrid” defects. Defects formed in a transition breaking a global symmetry tend to have energy density distributed throughout a region, whereas those formed by gauge symmetry breaking are more localized. Hybrid defects are produced by a series of phase transitions, leaving an interacting network of defects of different kinds. See, e.g., Vilenkin and Shellard (2000), for further discussion.

⁴⁵However, the sense in which the two theories are scale invariant is different; see, e.g., § 5.1.1 of Martin and Brandenberger (2001). Many defect models are scale invariant only over a limited dynamical range; for example, in models of defect formation via strings scale invariance is broken at the matter-radiation transition.

the interactions among the defects and their gravitational effects. The problem is exceedingly difficult because the evolution of defects is nonlinear and researchers have relied primarily on numerical simulations. Physically plausible suggestions regarding evolution of defects have often been undercut by numerical work. Throughout the 1980s, for example, the general picture of how strings seeded galaxy formation changed considerably in light of numerical simulations establishing details regarding the size of typical closed loops of strings and the behavior of open strings.⁴⁶ These two problems are exacerbated by uncertainty regarding the relevant fundamental physics. The details of the phase transitions depend on specific features of the physics—for example, the vacuum manifold is fixed by the full symmetry group G and its unbroken subgroup H , but these differ among proposed extensions of the standard model.

Despite these difficulties, by around 1997, there was a consensus regarding the generic consequences of structure formation via defects and the contrast with the consequences of inflation.⁴⁷ Structure formation via topological defects is “active” in the sense that the network of defects persists over time and continues to interact gravitationally with the other constituents. More precisely, in the evolution equation for perturbations of the cosmological model, there is a source term, representing the stress-energy of the network of defects. Determining the evolution of the perturbations thus requires calculating the evolution of this source term, based on the nonlinear dynamics of the network of defects. (For example, in the case of cosmic strings, the nonlinear dynamics describes the growth of the network of strings with the cosmological expansion and the different kinds of interactions among strings.) Perturbations produced in defect theories “decohere” (as first noted by Albrecht et al. (1996)) in the sense that fluctuations at all wave numbers are not in phase. This is a consequence of the nonlinear evolution of the source term, which leads to mixing of perturbations across different modes. The perturbations are also non-Gaussian due to the correlations that this mixing produces between perturbations. Finally, defects generate scalar, vector, and tensor perturbations of roughly equal magnitude.

This account of structure formation contrasts sharply with that based on inflation. Despite debates regarding how inflation related to particle physics, consensus was achieved regarding the consequences of inflation for structure formation.⁴⁸ Consider a massless, minimally coupled scalar field ϕ evolving in a background FLRW model. Due to the symmetry of the FLRW models, the Fourier modes ϕ_k of ϕ are uncoupled, and each mode evolves during slow-roll inflation according to the

⁴⁶See Vilenkin and Shellard (2000, Chapter 11) for an overview; the closing section (p. 342) emphasizes the changes in the account due to numerical simulations of the evolution of string networks.

⁴⁷Several groups published calculations at around this time supporting the general picture I summarize here; see, e.g., Magueijo et al. (1996) and Pen et al. (1997). See Durrer et al. (2002) for a comprehensive review of this area with further references to the original papers and Brandenberger (1994) for an earlier review.

⁴⁸Mukhanov et al. (1992) is the canonical review article regarding structure formation.

equation of a damped harmonic oscillator.⁴⁹ For modes such that $\frac{k}{R} \ll H$, the damping term is negligible, whereas those with $\frac{k}{R} \gg H$ will evolve like an overdamped oscillator and “freeze in” with a fixed amplitude. The inflationary account runs very roughly as follows. All the modes ϕ_k are assumed to be in their ground state prior to inflation. For $\frac{k}{R} \ll H$, the modes evolve adiabatically, remaining in their ground states, given that Equation (9.8) is approximately the equation for a harmonic oscillator. This account is not sensitive to exactly when a given mode is assumed to be “born” in its ground state. During inflation the modes scale with the exponential expansion whereas H is approximately constant. Due to this scaling behavior, modes will reach the horizon scale $\frac{k}{R} \approx H$ —“horizon exit.” The damping term is no longer negligible and the modes “freeze in” as they cross the horizon. Modes then “reenter” the horizon later given that the Hubble radius grows more rapidly than the modes after the inflationary stage has ended. Finally, these modes are treated as classical density perturbations upon re-entering the horizon.⁵⁰ This leads to a nearly scale-invariant spectrum; it is not *exactly* scale invariant because the Hubble radius is not exactly constant throughout inflation. The amplitude of the perturbations that are frozen in at horizon exit depends upon the details of the particular inflationary model under consideration.

The inflationary account differs in a number of respects from structure formation via defects. Inflation is a “passive” account of structure formation: there is no source term in the evolution equation, and in the linear regime, the solution is fixed by the initial conditions. Roughly speaking, in inflation the perturbations evolve “on their own” after being imprinted at early times, whereas in the defect theories, the network of defects persists and continues to seed structure formation. The most striking contrast is that inflation leads to phase coherence of the perturbations, because the dynamics described above leads to synchronization of the Fourier modes. Generically inflation predicts an oscillatory pattern in the angular power spectrum of temperature fluctuations in the CMBR, known as Doppler peaks.⁵¹

These contrasts between defect formation and inflation lead to quite strikingly different predictions for what should be observed in the CMBR (as illustrated in Figure 9.3). In addition, theorists working on these calculations through the 1990s were correct to expect that further observations of the CMBR would be able to

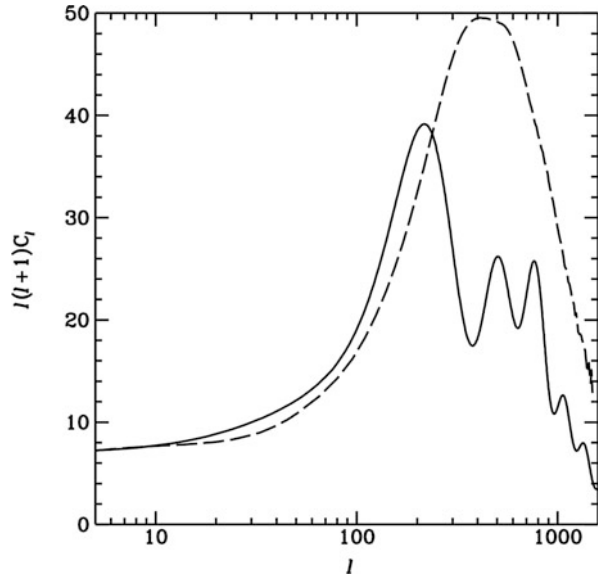
⁴⁹The equation can be derived from the action for the scalar field minimally coupled to gravity (with various simplifications, such as neglecting metric perturbations):

$$\frac{d^2\phi_k}{dt^2} + 3H\frac{d\phi_k}{dt} + \frac{k^2}{R^2}\phi_k = 0. \quad (9.8)$$

⁵⁰Although I do not have space to discuss the issue further here, this step involves a quantum to classical transition.

⁵¹The angular power spectrum characterizes the variations in temperature of the CMBR, i.e., the amount of temperature variation across different points of the sky versus the angular frequency ℓ . Small values of ℓ correspond to temperature variations with a large angular scale. See, e.g., Liddle and Lyth (2000, § 5.2), for further discussion of the angular power spectrum.

Fig. 9.3 This figure (from Albrecht et al. 1996) shows the predicted angular power spectrum of temperature fluctuations in the CMBR from a particular model of cosmic strings (dashed line) and a generic inflationary model (solid line).



discriminate among competing accounts. In particular, defect theories fail to predict strong secondary oscillations evident in subsequent CMBR observations. These features are “washed out” due to decoherence, whereas in inflationary accounts there are coherent standing wave oscillations in the baryon density that lead to strong secondary peaks. The position of the first peak also differs in inflation and topological defect models, with defect models generally predicting a primary peak at a larger multipole moment ($\ell \geq 300$) than inflation ($\ell \approx 200$). Observational results starting in the late 1990s and culminating in the WMAP results (3-year results published in 2003) provided decisive support for inflation with respect to both of these features.

In addition to the physical contrast between the mechanisms for structure formation, there are important methodological contrasts between the two approaches. First, despite uncertainty regarding the detailed physics of the phase transitions, the account of structure formation via defects is constrained enough by general theoretical principles to produce specific observational signatures. Physicists working on defects often highlighted this rigidity as a virtue of the theory, characterizing it as “falsifiable” in a Popperian sense. Second, accounts based on topological defects do not address the problems related to initial conditions highlighted by Guth. In effect, the theory starts from the same initial conditions as the standard FLRW models, with the exception that the initial seed perturbations were produced dynamically rather than fixed by hand. This methodological contrast did not play a role in the detailed evaluation of structure formation via topological defects in light of CMBR observations. Those who accepted Guth’s approach to fine-tuning and initial conditions could still use defects, however. Inflation could still be invoked to solve

the problems related to initial conditions (see, e.g., Vilenkin and Shellard 2000), as long as inflation set the stage for a subsequent phase transition that would produce appropriate topological defects.

The sharp contrast described above is based on assuming that there is only one primary mechanism for the formation of primordial perturbations. Yet the current theory of early universe phase transitions does not enforce such exclusivity. Ruling out models combining inflation and topological defects makes for a clearer theoretical contrast, amenable to decisive observational tests. But, as far as I am aware, there is little evidence that nature's choices are so conveniently circumscribed.

9.5 Characterizing Empirical Success

The fate of topological defects illustrates the power of contemporary cosmological observations. Within the last 50 years, cosmology has gone from being a field with only “2 1/2 facts”⁵² to a field with data that is sufficiently rich to warrant conclusions regarding novel physics far beyond the reach of earthbound accelerators. By the turn of the millennium, structure formation via topological defects was not only falsifiable but apparently falsified by cosmological observations.⁵³ Inflation did not share this fate; it is clearly compatible with the CMBR observations that ruled out defects. But in what sense does current observational data support inflationary theory? How should we characterize the empirical success or predictive power of the theory?

Debates in the physics literature have typically framed this question in terms of falsifiability. Has inflation avoided falsification just because it is unfalsifiable? Consider, for example, whether inflation could be falsified by finding that $\Omega_0 \approx 1$.⁵⁴ Flatness is often cited as an unambiguous, correct prediction of inflation. Guth (1997a), for example, emphasizes the extraordinary precision of the inflationary prediction—a correct value of Ω at the end of inflation to 15 significant figures! There are two reasons, however, to doubt that this is so straightforward.⁵⁵ First,

⁵²Peter Scheuer made this remark in the course of warning a student, Malcolm Longair, about the current status of cosmology in 1963; the list included (1) that the sky is dark at night, (2) that the galaxies recede, and (2 1/2) that the universe is evolving (qualified as a half fact due to its uncertainty).

⁵³Observations seem to rule out topological defects as the primary mechanism for generating large-scale structure. However, defects might still play a role as part of the full account of the formation of structure or in other aspects of early universe cosmology, such as baryogenesis.

⁵⁴The falsifiability of inflation, focusing in part on flatness, is addressed quite directly in a number of papers in Turok (1997), in particular the contributions by Linde, Steinhardt, Guth, and Albrecht. This has been a perennial subject of debate since the early days of inflation.

⁵⁵The question was particularly pressing throughout the 1990s, when the evidence seemed to favor open cosmological models with $\Omega_0 \approx 0.2$ – 0.3 , although there was not a general consensus. See, e.g., Coles and Ellis (1997) for a detailed argument in favor of an open universe. However, the consensus had begun to shift in favor of a flat universe by 1998. Peebles and David Schramm were

for *any* particular value of Ω_0 , there is a corresponding “initial” value $\Omega(t_p)$, whether inflation occurred or not. Thus the prediction has to be regarded as a probabilistic claim: for “highly probable” or “reasonable” initial conditions inflation yields $\Omega_0 = 1$. But, as discussed in Section 9.3.1 above, it is not clear what to make of these probabilistic claims without a measure over the space of initial values of Ω . The second objection is that the inflationary paradigm is too flexible to yield falsifiable predictions. In the mid-1990s, theorists constructed “open models” of inflation that yield a lower value of Ω_0 (see, e.g., Bucher et al. 1995). At most one might claim that a subset of inflationary models could be ruled out by finding $\Omega_0 \approx 1$, with further disagreement over whether this subset includes all of the “natural” or “reasonable” models of inflation. Rather than an unambiguous, falsifiable prediction, we are left with equivocal judgments regarding the probability assigned to initial conditions and the plausibility of different inflationary models.

Discussions of the falsifiability of inflation often draw Liddle and Lyth’s distinction quoted in the introduction between inflation “as a theory of initial conditions” and inflation as a theory of structure formation.⁵⁶ The account of structure formation appears to have definitive, falsifiable consequences. Several observational signatures—Gaussianity, near scale invariance—follow directly from the description of the dynamical evolution of the modes of a quantum field through horizon crossing. This dynamical mechanism for generating perturbations is a direct consequence of the defining feature shared by all inflationary models, given that it depends on the evolution of the Hubble constant during exponential expansion. Thus one might hope to avoid the above objections: the production of density perturbations is independent of assumptions regarding initial conditions, and the account is generic in the sense of being common to all models of inflation. But does the success of inflation simply exploit the malleability of the “inflaton” field and its potential? Note, for example, that the amplitude of the density perturbations needed for accounts of structure formation is used to constrain the parameters of the inflaton field. For this reason, Peebles (1999b) classifies the amplitude of the density perturbations as a “diagnostic” rather than a successful prediction.

Hollands and Wald (2002) argued that there is not such a clear contrast in terms of initial conditions. In particular, the inflationary account of the dynamical evolution of the modes of a quantum field through horizon crossing assumes that the modes are initially in their ground state. This is a plausible assumption given that the modes with cosmologically significant length scales will be well inside the Hubble radius prior to the inflationary phase. Since the modes evolve adiabatically before horizon

invited to convene a “great debate” on the issue in April of 1998. Due to Schramm’s death, the debate was rescheduled for October of 1998, with Michael Turner taking Schramm’s place. But given that Peebles and Turner both agreed that the evidence decisively favored a flat universe, they changed the subject of the debate to “Is Cosmology Solved?” (Peebles 1999a; Turner 1999).

⁵⁶The distinction is perhaps too quick, given that there are some predictions related to initial conditions. For example, inflation predicts that the observed universe is topologically simply connected; inflation is incompatible with compact topology at sub-horizon scales. Evidence that the universe is multiply connected would rule out inflation.

crossing, the exact time at which they are taken to be “born” in their ground state is unimportant. Hollands and Wald (2002) construct a simple model that produces a similar spectrum of density perturbations *without an inflationary phase* based on a different *Ansatz* for the initial conditions for these modes. Their model describes quantized sound waves in a perfect fluid, with the same “overdamping” of modes with $\lambda \gg H^{-1}$ as in inflation. By contrast with inflation, there is no horizon crossing, so it is significant precisely when the modes are taken to be in a vacuum state. Hollands and Wald (2002) propose to take the modes to be “born” in a ground state when their proper wavelength is equal to the Planck scale, motivated by considerations of the domain of applicability of semiclassical quantum gravity.⁵⁷ This hypothesis combined with the dynamics governing the evolution of the modes leads to a scale-invariant perturbation spectrum. The significance of this result for present purposes is that it undermines claims that the theory of structure formation does not depend on arguments regarding plausible initial conditions.

Stepping back from the details of inflation for a moment, it should be clear that there are important questions regarding both how to characterize a theory’s empirical success and what a given degree of success establishes. It is unfortunate that these questions are still treated in the physics literature in terms of “falsifiability,” and I will briefly sketch an alternative drawing on recent studies of Newton’s methodology (Harper 2002; Smith 2002). On this approach, empirical success is defined in terms of the ability to determine consistent values of theoretical parameters from multiple, independent bodies of data. Consider, for example, Newton’s argument in favor of a universal force of gravity in the *Principia*. Newton takes the theoretical framework provided by the laws of motion to be exact, and the array of mathematical results applying to forces in general then allows him to infer properties of the gravitational force from the observed motions of the planets, their satellites, and various other bodies (such as pendulums). The famous precession theorem is a particularly beautiful example: Newton shows that for approximately circular orbits, the motion of the apsides measures the exponent of the power law.⁵⁸ Taking the exponent in the power law for gravity as our example of a theoretical parameter, there are several lines of argument from diverse, independent bodies of data that fix the value as very close to -2 . This account acknowledges that the theory

⁵⁷The modes will be “born” at different times, continually “emerging out of the spacetime foam” (or whatever description the full theory of quantum gravity provides), with the modes relevant to large-scale structure born at times much earlier than the Planck time. By way of contrast, in the usual approach the modes at all length scales are specified to be in a ground state at a particular time, such as the Planck time. But the precise time at which one stipulates the field modes to be in a vacuum state does not matter given that the sub-horizon modes evolve adiabatically.

⁵⁸The apsidal angle θ is the angle through which the radius vector rotates between two consecutive apsides, which are points on the orbit of maximum (aphelion) or minimum (perihelion) distance from the force center. Newton establishes (Book I, Proposition 45) that for approximately circular orbits under a centripetal force varying as $f \propto r^{n-3}$, the apsidal angle is given by $n = \left(\frac{\theta}{\pi}\right)^2$. For stable orbits, the radius vector rotates through π between the aphelion and perihelion, such that $n = 1$ and $f \propto r^{-2}$; and for nearly stable orbits, the force is approximately $f \propto r^{-2}$.

requires some data as “input” to enable further predictions. Other bodies of data that can be used to constrain the same parameter value then provide independent checks. Harper (1990, 2007) argues that Newtonian characterization of empirical success is much more demanding than mere predictive accuracy. A theory that achieves predictive accuracy by “curve fitting” (exploiting theoretical flexibility) will suffer by comparison with a more rigid theory on the Newtonian account.

The strength of Newton’s empirical argument for universal gravitation bears directly on two potential objections. First, why should one accept gravity as a “real force” given that it apparently involved action at a distance? Although the issue is complicated, Newton clearly held that the empirical case was sufficient to establish the reality of gravitational force despite uncertainty regarding its underlying cause and certainty that it is not a “mechanical” cause (i.e., due to contact action). Second, from a modern perspective, why should Newton’s theory be preserved as a limiting case of general relativity? If we regarded the theory merely as a predictively accurate curve fit, rather than an accurate systematic treatment of physical relationships within a limited domain, there would be no reason to expect general relativity to recover anything more than the predictions themselves. Speaking more generally, the first kind of objection relates to unresolved problems. In some cases the empirical success of a theory is sufficient to warrant acceptance even in light of open physical questions. The second challenge regards the use of a theory as a step toward further theories. Sufficient empirical success warrants preserving not just the predictions of the theory but the physical relationships it ascribes to systems within its domain.

Returning to the case of inflation, there are two similar challenges. First, there are various open problems regarding the place of an “inflaton” field within particle physics at the appropriate energy scales and the coupling of a scalar field to gravity. The cosmological constant problem is sometimes characterized as the Achilles heel of inflation. Inflation is built on the assumption that the false vacuum energy of the inflaton field couples to gravitation. But if this is so, the vacuum energy density of other quantum fields should contribute to gravity as an effective cosmological constant. A comparison between the vacuum energy density calculated in QFT and observational limits on the cosmological constant in GR reveals an incredible discrepancy of some 120 orders of magnitude! As Frank Wilczek commented in a review of the Nuffield workshop:

It is surely an act of cosmic *chutzpah* to use this dismal theoretical failure [in understanding the cosmological constant] as a base for erecting theoretical superstructures, but of course this is exactly what is done in current inflationary models. (Hawking et al. 1983, 476, original emphasis)

Second, cosmologists have often suggested that the requirement to find an inflaton field should serve as a constraint on particle physics. This is certainly appealing, as a successful case for inflation would provide a strong constraint at energy scales with few observational constraints from earthbound accelerators.

On this approach, the question to ask regarding inflation is not whether it makes various “falsifiable” predictions but to what extent do the observational data allow

us to infer the details of inflation? On the assumption that inflation is correct, what do the data allow us to infer about the inflaton field and its effective potential $V(\phi)$? In these terms the account of inflation as a theory of structure formation provides a richer set of constraints on the theory. The solution of the horizon and flatness problems constrains the duration of the inflationary phase: the pre-inflationary patch has to grow larger than the observed universe, at a minimum. The inflationary stage will last sufficiently long if the potential $V(\phi)$ is suitably flat and satisfies the “slow-roll” conditions described in Section 9.3.2 above. The account of structure formation, by contrast, provides more detailed constraints. The fluctuation modes that seed the formation of structure depend on the properties of the effective potential $V(\phi)$ at the time when they cross the horizon. (There is a limit on the part of the potential that can be constrained in this way, given that only some of the modes will have reentered the horizon as observable density perturbations.) This opens up the prospect of reconstructing the inflaton potential based on observations of the CMBR. Whether the reconstruction provides sufficient empirical warrant to answer the challenges above is another question.

References

- Albrecht, A., Coulson, D., Ferreira, P., & Magueijo, J. (1996). Causality, randomness, and the microwave background. *Physical Review Letters*, 76(9), 1413.
- Albrecht, A., & Steinhardt, P. (1982). Cosmology for grand unified theories with induced symmetry breaking. *Physical Review Letters*, 48, 1220–1223.
- Barbour, J. B., & Pfister, H. (1995). *Mach’s principle: From Newton’s bucket to quantum gravity* (Vol. 6). New York: Springer.
- Bardeen, J. M. (1980). Gauge invariant cosmological perturbations. *Physical Review D*, 22, 1882–1905.
- Bardeen, J. M., Steinhardt, P. J., & Turner, M. S. (1983). Spontaneous creation of almost scale-free density perturbations in an inflationary universe. *Physical Review D*, 28, 679.
- Barrow, J. D. (1980). Galaxy formation - the first million years. *Royal Society of London Philosophical Transactions A*, 296, 273–288.
- Barrow, J. D., & Turner, M. S. (1981). Inflation in the universe. *Nature*, 292, 35–38.
- Barrow, J. D., & Turner, M. S. (1982). The inflationary universe—birth, death, and transfiguration. *Nature*, 298, 801–805.
- Blau, S. K., & Guth, A. (1987). Inflationary cosmology. In Hawking, S. W. & Israel, W. (Eds.), *300 years of gravitation* (pp. 524–603). Cambridge: Cambridge University Press.
- Blewitt, G., LoSecco, J. M., Bionla, R. M., Bratton, C. B., Casper, D., & Chrysiopoulou, P. (1985). Experimental limits on the free proton lifetime for two and three-body decay modes. *Physical Review Letters*, 55, 2114–2117.
- Bonnor, W. B. (1956). The formation of the nebulae. *Zeitschrift für Astrophysik*, 39, 143–159.
- Brandenberger, R. H. (1994). Topological defects and structure formation. *International Journal of Modern Physics A*, 9(13), 2117–2189.
- Brawer, R. (1996). *Inflationary cosmology and the horizon and flatness problems: The mutual constitution of explanation and questions*. Master’s thesis, Massachusetts Institute of Technology, Department of Physics.
- Bucher, M., Goldhaber, A. S., & Turok, N. (1995). An open universe from inflation. *Physical Review D*, 52, 3314–3337.

- Coleman, S. (1985). *Aspects of symmetry: Selected Erice lectures*. Cambridge: Cambridge University Press.
- Coles, P., & Ellis, G. F. R. (1997). *Is the universe open or closed?* Cambridge: Cambridge University Press.
- Collins, C. B., & Stewart, J. M. (1971). Qualitative cosmology. *Monthly Notices of the Royal Astronomical Society*, 153, 419–434.
- Dicke, R., & Peebles, P. J. E. (1979). The big bang cosmology—enigmas and nostrums. In Hawking, S. W. & Israel, W. (Eds.), *General relativity: An Einstein centenary survey* (pp. 504–517). Cambridge: Cambridge University Press.
- Dicke, R. H. (1969). *Gravitation and the universe: Jayne lectures for 1969*. Philadelphia: American Philosophical Society.
- Durrer, R., Kunz, M., & Melchiorri, A. (2002). Cosmic structure formation with topological defects. *Physics Reports*, 364(1), 1–81.
- Earman, J., & Mosterin, J. (1999). A critical analysis of inflationary cosmology. *Philosophy of Science*, 66(1), 1–49.
- Ellis, G. F. R., & Madsen, M. S. (1988). The evolution of Ω in inflationary universes. *Monthly Notices of the Royal Astronomical Society*, 234, 67–77.
- Ellis, G. F. R., & Rothman (1993). Lost horizons. *American Journal of Physics*, 61(10), 883–893.
- Gamow, G. (1952). The role of turbulence in the evolution of the universe. *Physical Review*, 86, 251.
- Gamow, G. (1954). On the formation of protogalaxies in the turbulent primordial gas. *Proceedings of the National Academy of Science*, 40, 480–484.
- Gamow, G., & Teller, E. (1939). On the origin of great nebulae. *Physical Review*, 55(7), 654.
- Guth, A. (1981). Inflationary universe: A possible solution for the horizon and flatness problems. *Physical Review D*, 23, 347–356.
- Guth, A. (1997a). *The inflationary universe*. Reading, MA: Addison-Wesley.
- Guth, A. (1997b). Thesis: Inflation provides a compelling explanation for why the universe is so large, so flat, and so old, as well as a (almost) predictive theory of density perturbations. In Turok, N. (Ed.), *Critical dialogues in cosmology* (pp. 233–248). Singapore: World Scientific.
- Guth, A., & Tye, S.-H. H. (1980). Phase transitions and magnetic monopole production in the very early universe. *Physical Review Letters*, 44, 631–634.
- Guth, A. H., & Pi, S. Y. (1982). Fluctuations in the new inflationary universe. *Physical Review Letters*, 49, 1110–1113.
- Harper, W. (1990). Newton's classic deductions from phenomena. *Proceedings of the 1990 Biennial Meeting of the Philosophy of Science Association*, 2, 183–196.
- Harper, W. (2002). Newton's argument for universal gravitation. In Cohen, I. B. & Smith, G. E. (Eds.), *Cambridge companion to Newton* (pp. 174–201). Cambridge: Cambridge University Press.
- Harper, W. (2007). Newton's methodology and Mercury's perihelion before and after Einstein. *Philosophy of Science*, 74, 932–942.
- Harrison, E. R. (1967a). Normal modes of vibrations of the universe. *Reviews of Modern Physics*, 39, 862–882.
- Harrison, E. R. (1967b). On the origin of structure in certain models of the universe. Introductory report. In *Liege international astrophysical colloquia* (Vol. 14, p. 15).
- Harrison, E. R. (1968). On the origin of galaxies. *Monthly Notices of the Royal Astronomical Society*, 141, 397–407.
- Harrison, E. R. (1970). Fluctuations at the threshold of classical cosmology. *Physical Review D*, 1, 2726–2730.
- Hawking, S. W. (1982). The development of irregularities in a single bubble inflationary universe. *Physics Letters B*, 115, 295–297.
- Hawking, S. W., Gibbons, G. W., & Siklos, S. T. C. (Eds.). (1983). *The very early universe*. Cambridge: Cambridge University Press.
- Hawking, S. W., & Moss, I. G. (1982). Supercooled phase transitions in the very early universe. *Physics Letters B*, 110, 35–38.

- Hollands, S., & Wald, R. (2002). Essay: an alternative to inflation. *General Relativity and Gravitation*, 34, 2043–2055.
- Ijjas, A., Steinhardt, P. J., & Loeb, A. (2013). Inflationary paradigm in trouble after planck2013. *Physics Letters B*, 723(4), 261–266.
- Jeans, J. H. (1902). The stability of a spherical nebula. *Philosophical Transactions of the Royal Society A*, 199, 1–53.
- Kibble, T. W. B. (1976). Topology of cosmic domains and strings. *Journal of Physics A*, 9, 1387–1397. Reprinted in Bernstein and Feinberg (1986).
- Kirzhnits, D. A. (1972). Weinberg model in the hot universe. *JETP Letters*, 15, 529–531.
- Kolb, E. W., & Turner, M. S. (1990). *The early universe. Vol. 69. Frontiers in physics*. New York: Addison-Wesley.
- Kragh, H. (1996). *Cosmology and controversy*. Princeton: Princeton University Press.
- Lemaître, G. (1933). L'univers en expansion. *Annales de la Société Scientifique de Bruxelles*, 53, 51–85.
- Liddle, A., & Lyth, D. (2000). *Cosmological inflation and large-scale structure*. Cambridge: Cambridge University Press.
- Lifshitz, Y. M. (1946). On the gravitational stability of the expanding universe. *Journal of Physics USSR*, 10, 116–129.
- Lightman, A., & Brawer, R. (1990). *Origins: The lives and worlds of modern cosmologists*. Cambridge: Harvard University Press.
- Linde, A. (1979). Phase transitions in gauge theories and cosmology. *Reports on Progress in Physics*, 42, 389–437.
- Linde, A. (1982). A new inflationary universe scenario: a possible solution of the horizon, flatness, homogeneity, isotropy, and primordial monopole problems. *Physics Letters B*, 108, 389–393.
- Linde, A. (1990). *Particle physics and inflationary cosmology*. Amsterdam: Harwood Academic Publishers.
- Longair, M. (2006). *The cosmic century: A history of astrophysics and cosmology*. Cambridge: Cambridge University Press.
- Longair, M. (2007). *Galaxy formation*. New York: Springer.
- Madsen, M. S., & Ellis, G. F. R. (1988). The evolution of ω in inflationary universes. *Monthly Notices of the Royal Astronomical Society*, 234, 67–77.
- Magueijo, J., Albrecht, A., Coulson, D., & Ferreira, P. (1996). Doppler peaks from active perturbations. *Physical Review Letters*, 76(15), 2617.
- Markov, M. A., & West, P. C. (Eds.). (1984). *Quantum gravity. In Proceedings of the Second Seminar on Quantum Gravity; Moscow, October 13–15, 1981*. New York: Plenum Press.
- Martin, J., & Brandenberger, R. H. (2001). The trans-Planckian problem of inflationary cosmology. *Physical Review D*, 63, 123501.
- Misner, C. W. (1969). Mixmaster universe. *Physical Review Letters*, 22, 1071–1074.
- Mukhanov, V. F., & Chibisov, G. V. (1981). Quantum fluctuations and a nonsingular universe. *JETP Letters*, 33, 532–535.
- Mukhanov, V. F., Feldman, H. A., & Brandenberger, R. H. (1992). Theory of cosmological perturbations. Part 1: Classical perturbations. Part 2: Quantum theory of perturbations. Part 3: Extensions. *Physics Reports*, 215, 203–333.
- Nanopoulos, D. V., Olive, K. A., & Srednicki, M. (1983). After primordial inflation. *Physics Letters B*, 127, 30–34.
- Olive, K. A. (1990). Inflation. *Physics Reports*, 190, 307–403.
- Pagels, H. R. (1984). New particles and cosmology. In *Eleventh Texas Symposium on Relativistic Astrophysics* (p. 15). New York: Academy of Sciences.
- Partridge, R. B. (1980). New limits on small-scale angular fluctuations in the cosmic microwave background. *The Astrophysical Journal*, 235, 681–687.
- Peacock, J. R. (1999). *Cosmological physics*. Cambridge: Cambridge University Press.
- Peebles, P. J. E. (1965). The black-body radiation content of the universe and the formation of galaxies. *Astrophysical Journal*, 142, 1317.

- Peebles, P. J. E. (1967). The gravitational instability of the universe. *Astrophysical Journal*, 147, 859.
- Peebles, P. J. E. (1968). Formation of galaxies in classical cosmology. *Nature*, 220, 237.
- Peebles, P. J. E. (1971). *Physical cosmology*. Princeton: Princeton University Press.
- Peebles, P. J. E. (1980). *Large-scale structure of the universe*. Princeton: Princeton University Press.
- Peebles, P. J. E. (1982). Large-scale background temperature and mass fluctuations due to scale-invariant primeval perturbations. *Astrophysical Journal*, 263, L1–L5.
- Peebles, P. J. E. (1999a). Is cosmology solved? An astrophysical cosmologist's viewpoint. *Publications of the Astronomical Society of the Pacific*, 111, 274–284.
- Peebles, P. J. E. (1999b). Summary: Inflation and traditions of research. Arxiv preprint astro-ph/9905390.
- Peebles, P. J. E., & Yu, J. T. (1970). Primeval adiabatic perturbation in an expanding universe. *The Astrophysical Journal*, 162, 815–836.
- Pen, U.-L., Seljak, U., & Turok, N. (1997). Power spectra in global defect theories of cosmic structure formation. *Physical Review Letters*, 79(9), 1611.
- Penrose, R. (1986). Review of the very early universe. *The Observatory*, 106, 20–21.
- Penrose, R. (1989). Difficulties with inflationary cosmology. *Annals of the New York Academy of Sciences*, 271, 249–264.
- Penrose, R. (2004). *The road to reality*. London: Jonathan Cape.
- Press, W. H., & Schechter, P. (1974). Formation of galaxies and clusters of galaxies by self-similar gravitational condensation. *The Astrophysical Journal*, 187, 425–438.
- Press, W. H., & Vishniac, E. T. (1980). Tenacious myths about cosmological perturbations larger than the horizon size. *Astrophysical Journal*, 239, 1–11.
- Rindler, W. (1956). Visual horizons in world models. *Monthly Notices of the Royal Astronomical Society*, 116, 662–677.
- Sakharov, A. D. (1966). The initial state of an expanding universe and the appearance of a nonuniform distribution of matter. *Soviet Physics JETP*, 22, 241–249, Reprinted in *Collected Scientific Works*.
- Sandage, A. (1970). Cosmology: A search for two numbers. *Physics Today*, 23(2), 33–43.
- Shafi, Q., & Vilenkin, A. (1984). Inflation with SU(5). *Physical Review Letters*, 52, 691–694.
- Shellard, P. (2003). The future of cosmology: Observational and computational prospects. In G. Gibbons, E. Shellard & S. Rankin (Eds.), *The future of theoretical physics and cosmology* (pp. 755–780). Cambridge: Cambridge University Press.
- Smeenk, C. (2005). False vacuum: Early universe cosmology and the development of inflation. In A.J. Kox & J. Eisenstaedt (Eds.), *The universe of general relativity. Vol. 11. Einstein studies* (pp. 223–257). Boston: Birkhäuser.
- Smith, G. E. (2002). The methodology of the *Principia*. In I. B. Cohen & G. E. Smith (Eds.), *Cambridge companion to Newton* (pp. 138–173). Cambridge: Cambridge University Press.
- Starobinsky, A. (1978). On a nonsingular isotropic cosmological model. *Soviet Astronomy Letters*, 4, 82–84.
- Starobinsky, A. (1979). Spectrum of relic gravitational radiation and the early state of the universe. *JETP Letters*, 30, 682–685.
- Starobinsky, A. (1982). Dynamics of phase transitions in the new inflationary scenario and generation of perturbations. *Physics Letters B*, 117, 175–178.
- Starobinsky, A. (1983). The perturbation spectrum evolving from a nonsingular initially de-sitter cosmology and the microwave background anisotropy. *Soviet Astronomy Letters*, 9, 302–304.
- Steinhardt, P. (2002). Interview with Paul Steinhardt conducted by Chris Smeenk (100 pp.), manuscript, to be deposited in the Oral History Archives at the American Institute of Physics.
- Steinhardt, P. J., & Turner, M. S. (1984). A prescription for successful new inflation. *Physical Review D*, 29, 2162–2171.
- Turner, M. (1999). Cosmology solved? Quite possibly! *Publications of the Astronomical Society of the Pacific*, 111, 264–273.

- Turok, N. (Ed.) (1997). *Critical dialogues in cosmology*. Singapore: World Scientific.
- Unruh, W. G. (1997). Is inflation the answer? In N. Turok (Ed.), *Critical dialogues in cosmology* (pp. 249–264). Singapore: World Scientific.
- Vachaspati, T., & Trodden, M. (1999). Causality and cosmic inflation. *Physical Review D*, 61(2), 23502.
- Vilenkin, A., & Shellard, E. (2000). *Cosmic strings and other topological defects*. Cambridge: Cambridge University Press.
- Weinberg, S. (1972). *Gravitation and cosmology*. New York: Wiley.
- Weinberg, S. (2008). *Cosmology*. Oxford: Oxford University Press.
- Zeldovich, Y. B. (1965). Survey of modern cosmology. *Advances in Astronomy and Astrophysics*, 3, 241–391.
- Zel'dovich, Y. B. (1972). A hypothesis, unifying the structure and the entropy of the universe. *Monthly Notices of the Royal Astronomical Society*, 160, 1–3.
- Zel'dovich, Y. B., & Khlopov, M. Y. (1978). On the concentration of relic magnetic monopoles in the universe. *Physics Letters B*, 79, 239–241.
- Zel'dovich, Y. B., Kobzarev, I. Y., & Okun, L. B. (1975). Cosmological consequences of a spontaneous breakdown of a discrete symmetry. *Soviet Physics JETP*, 40, 1–5.
- Zel'dovich, Y. B., & Novikov, I. (1983). In G. Steigman, K. Thorne, & W. Arnett (Eds.), *Relativistic astrophysics* (Two volumes; E. Arlock and L. Fishbone, Trans.). Chicago: University of Chicago Press.

Chapter 10

Problems with Modified Theories of Gravity, as Alternatives to Dark Energy



Norbert Straumann

10.1 Introduction

The phenomenologically very successful cosmological “concordance model,” within the framework of general relativity (GR), leaves us with the mystery of dark energy (DE). Since no satisfactory explanation of DE has emerged so far,¹ it is certainly reasonable to investigate whether possible modifications of GR might change the late expansion rate of the universe. After all, GR has not yet been tested on cosmological scales.

Modified gravity models have to be devised such that they pass the stringent solar system tests and are compatible with the rich body of cosmological data that support the concordance model (Λ CDM model). At the same time, the theories should be consistent on a fundamental level. Since we are dealing with higher spin equations, possible acausalities are, for instance, a serious issue.

Apart from all that, one should not forget that the old profound vacuum energy problem (Straumann 2007) and the cosmic coincidence problem remain, and thus extreme fine-tuning is unavoidable. This holds, of course, also for all dynamical models of DE (Copeland et al. 2006).

In my brief review, I shall mainly concentrate on the so-called $f(R)$ gravity. This is the simplest modification. Moreover, there have been some recent developments that I find interesting. After some generalities, many of you know very well, I shall discuss the weak-field limit and solar system tests. For some time there was a lot of confusion on this issue, with conflicting statements. We shall, however, see that the

¹See, e.g., Copeland et al. (2006) and Straumann (2007) and references therein.

N. Straumann (✉)

Physik-Institut, University of Zurich, Winterthurerstrasse 190, CH–8057 Zurich, Switzerland

weak field approximation may break down and a so-called Chameleon mechanism can be at work that hides a scalar degree of freedom of the theory on solar system scales.

There are $f(R)$ models that pass the solar system tests and are cosmologically almost indistinguishable from the successful Λ CDM model. Recently it was, however, discovered that these models are in serious trouble in the strong-field regime.

Some of the other modified gravity theories are even in greater difficulties. This will be briefly discussed in a final part.

10.2 Metric $f(R)$ Gravity

The simplest possibility of modifying GR is to replace the Einstein-Hilbert action $R - 2\Lambda$ of gravity by a nonlinear function $f(R)$ of the Ricci scalar R .² This introduces an additional scalar degree of freedom that can lead to an accelerated expansion of the universe at late times, induced by the Ricci scalar. One may call this “curvature DE” or “dark gravity.” The function f is quite arbitrary, and the theory loses, of course, a lot of predictive power. As many other people, I regard this class of modified gravity theories as instructive phenomenological toy models that change gravity in the *infrared* (large scales).

The variation of the gravitational action is

$$\delta \int f(R) \sqrt{-g} d^4x = \int \{R_{\alpha\beta} f'(R) - \frac{1}{2} g_{\alpha\beta} f(R) + g_{\alpha\beta} \nabla^2 f'(R) - \nabla_\alpha \nabla_\beta f'(R)\} \delta g^{\alpha\beta} \sqrt{-g} d^4x. \quad (10.1)$$

Diffeomorphism invariance implies that the tensor in the curly bracket has a vanishing covariant divergence. Therefore, the field equation

$$R_{\alpha\beta} f'(R) - \frac{1}{2} g_{\alpha\beta} f(R) + g_{\alpha\beta} \square f'(R) - \nabla_\alpha \nabla_\beta f'(R) = 8\pi G T_{\alpha\beta} \quad (10.2)$$

implies that the energy-momentum tensor $T^{\alpha\beta}$ is divergence-free. This implies, by a general result of Hawking, that matter propagates causally if $T^{\alpha\beta}$ satisfies the dominant energy condition. As expected from Lovelock’s theorem, the field equation is of fourth order in the metric if f is not a linear function.

²For an extensive review and literature, we refer to Sotiriou and Faraoni (2010).

It is easy to see that the de Sitter or anti-de Sitter metric, with $R_{\alpha\beta} = \Lambda g_{\alpha\beta}$, is a vacuum solution, if there is a constant Λ satisfying $f(4\Lambda) = 2\Lambda f'(4\Lambda)$.³ This indicates that the theory may naturally lead to cosmological acceleration. More on this later.

At first sight one may think that experience and insight from GR may not help us to get some understanding of what the complicated fourth-order field equations may describe. It is, however, known since long that $f(R)$ gravity models can be reformulated as scalar-tensor theories (Whitt 1984; Maeda 1989). To show this, we first rewrite the field equations in the following form ($\kappa^2 := 8\pi G$):

$$\begin{aligned} f'(R)G_{\alpha\beta} &= \kappa^2 T_{\alpha\beta} + \frac{1}{2}g_{\alpha\beta}[f(R) - Rf'(R)] \\ &\quad - g_{\alpha\beta}\square f'(R) + \nabla_\alpha \nabla_\beta f'(R). \end{aligned} \quad (10.3)$$

It is natural to introduce the scalar field $\phi := f'(R)$. We assume that $f'' \neq 0$, so that f' is at least locally invertible: $f' \circ \mathcal{R} = id$. Furthermore, let \mathcal{U} be the Legendre transform of f :

$$\mathcal{U}(\phi) = \mathcal{R}(\phi)\phi - f(\mathcal{R}(\phi)) \quad (10.4)$$

(thus $\mathcal{U}' = R$). With this we can rewrite (10.3) as

$$\phi G_{\alpha\beta} = \kappa^2 T_{\alpha\beta} - \frac{1}{2}g_{\alpha\beta} \mathcal{U}(\phi) - [g_{\alpha\beta}\square\phi - \nabla_\alpha \nabla_\beta \phi]. \quad (10.5)$$

This is just the Brans-Dicke equation with the Brans-Dicke parameter $\omega_{BD} = 0$ plus a potential term.⁴ This indicates that the weak-field limit may be in conflict with solar system tests, because these imply that the parameter ω_{BD} has to be very large: $\omega_{BD} > 40'000$. We shall see that this is indeed the case, but thanks to the potential term, there is an interesting way out.

Taking the trace of the original field equation (10.3), we obtain

$$3\square f'(R) + Rf'(R) - 2f(R) = \kappa^2 T. \quad (10.6)$$

In terms of the scalar field ϕ , this becomes

$$3\square\phi + 2\mathcal{U}(\phi) - \phi \mathcal{U}'(\phi) = \kappa^2 T. \quad (10.7)$$

³If this transcendental equation has a solution, any vacuum solution of GR with the corresponding Λ is obviously a vacuum solution of (10.2).

⁴Therefore, one expects that the Cauchy problem is well-posed. This is certainly the case for the vacuum theory, but with matter the problem is not completely settled; see Salgado (2006) and Salgado et al. (2008). In Salgado et al. (2008) two first-order strongly hyperbolic formulations of scalar-tensor theories are presented, which however do not include the exceptional case $\omega = -3/2$.

This nonlinear scalar field equation will play a crucial role. It shows that the scalar degree of freedom is truly dynamical. In contrast to GR, the scalar Ricci curvature does no more track the matter distribution.

We note at this point that the field equations (10.5) follow from the following action:

$$S = \frac{1}{2\kappa^2} \int [\phi R - \mathcal{U}(\phi)] \sqrt{-g} d^4x + S_M. \quad (10.8)$$

Since no kinetic energy for the ϕ -field appears in this action, one may be tempted to conclude that ϕ is not a dynamical field, but we have just seen that this is not the case.

For certain problems it can be useful to pass to a mathematically equivalent description by performing the conformal transformation (first studied by Pauli in letters to Jordan from 1953 Pauli 1985–99):

$$\tilde{g}_{\mu\nu} = \exp \left[\sqrt{\frac{2}{3}} \kappa \phi \right] g_{\mu\nu}, \quad \phi = \exp \left[\sqrt{\frac{2}{3}} \kappa \varphi \right]. \quad (10.9)$$

In terms of the new metric and the scalar field φ , called the *Einstein frame*, the action becomes

$$S_{EF} = \int \left[\frac{1}{2\kappa^2} R[\tilde{g}] - \frac{1}{2} \tilde{g}^{\alpha\beta} \partial_\alpha \varphi \partial_\beta \varphi - U(\varphi) \right] \sqrt{-\tilde{g}} d^4x + S_M[\tilde{g}_{\mu\nu} e^{-\beta\varphi}], \quad (10.10)$$

where $\beta := \sqrt{\frac{2}{3}} \kappa$, and

$$U(\varphi) = \mathcal{U}(\phi(\varphi))/2\kappa^2 \phi(\varphi)^2 = \frac{1}{2\kappa^2} e^{-2\beta\varphi} [e^{\beta\varphi} \mathcal{R}(e^{\beta\varphi}) - f(\mathcal{R}(e^{\beta\varphi}))]. \quad (10.11)$$

In contrast to the original *Jordan frame* description, the gravitational part of the action becomes the Einstein-Hilbert action of GR, but the coupling to matter is *nonminimal*. This implies that relative to the Levi-Civita connection belonging to the metric $\tilde{g}_{\mu\nu}$, the energy-stress tensor is no more conserved. In the Einstein frame, matter feels a new “fifth force” due to gradients of φ . While Newton’s constant is everywhere the same, the local particle physics thus varies. In the Jordan frame, on the other hand, the laws of physics in local inertial frames are universal, but the effective gravitational “constant” (G/ϕ) becomes spacetime dependent. Since the two descriptions are mathematically equivalent, observables are frame independent. It is then just a matter of convenience which one prefers to use. In what follows we will always work in the Jordan frame, except at one instance.

10.3 Generalized Friedmann Models

It is straightforward to derive the modified Friedmann equations. We consider only Friedmann-Lemaitre (-Robertson-Walker) spacetimes that are spatially flat. If $a(t)$ denotes the scale factor and $H = \dot{a}/a$ the Hubble rate, one finds

$$H^2 = \frac{K^2}{3f'}(\rho + \rho_{\text{eff}}), \quad (10.12)$$

$$\frac{\ddot{a}}{a} = -\frac{K^2}{6f'}[\rho + \rho_{\text{eff}} - 3(P + P_{\text{eff}})], \quad (10.13)$$

where ρ is the energy density and P the pressure⁵ Furthermore,

$$\rho_{\text{eff}} = \frac{1}{2}(Rf' - f) - 3H\dot{R}f'', \quad (10.14)$$

$$P_{\text{eff}} = \dot{R}^2 f''' + 2H\dot{R}f'' + \ddot{R}f'' + \frac{1}{2}(f - Rf') \quad (10.15)$$

are effective fluid contributions due to curvature (“curvature dark energy”). The sign of the corresponding effective equation of state parameter $w_{\text{eff}} := P_{\text{eff}}/\rho_{\text{eff}}$ is determined by that of P_{eff} since ρ_{eff} has to be nonnegative. Simple choices for $f(R)$ give strongly negative values for w_{eff} . For example, $f(R) = R - \mu^4/R$ gives $w_{\text{eff}} = -2/3$. As we noted before, the effective gravitational “constant” in (10.12) and (10.13) is G/f' and is thus R -dependent.

In general none of the standard energy conditions is satisfied for $f(R)$ models. In particular, $|P_{\text{eff}}|$ does not have to be smaller than ρ .

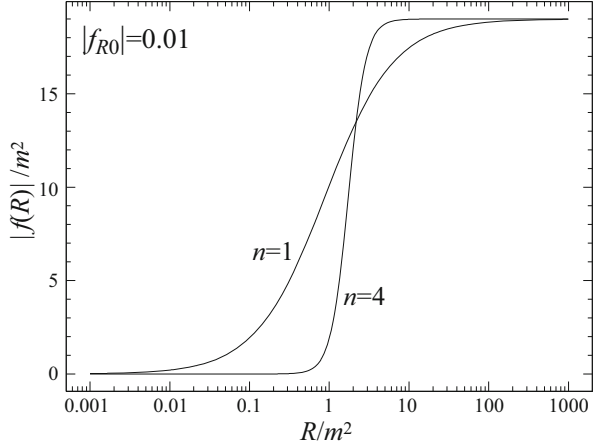
Clearly, the “energy conservation” $\dot{\rho} = -3H(\rho + P)$ follows, as in GR, from the field equations, i.e., from (10.12) and (10.13).

Since $\ddot{a}/a = \dot{H} + H^2$, we may regard the evolution equations (10.12) and (10.13) as a system of equations for H and R . This dynamical system has been extensively studied (see, e.g., Amendola et al. 2007). It is a mathematical fact⁶ that for any given expansion history $a(t)$, there exists a (nonunique) function f that reproduces this history by the corresponding $f(R)$ model. This does, however, not guarantee that the sequence of radiation-matter-acceleration eras is also reproduced. Indeed, the analysis in Fay et al. (2007) shows that *nonlinear* $f(R)$ models that reproduce, for example, exactly the history $a(t)$ of the Λ CDM model do not have the right sequence of cosmological eras with the required density parameters. However, it is possible to reproduce this sequence of eras if some deviations from the given history $H(z)$ are tolerated. An example for this, that is compatible with current observations, has been given by Hu and Sawicki (2007).

⁵For symmetry reasons $T^{\mu\nu}$ has the form of an ideal fluid of $T^{\mu\nu}$.

⁶For a dynamical system analysis of this reconstruction, see Fay et al. (2007).

Fig. 10.1 Plot of $|(f(R) - R)/m^2|$ from Hu and Sawicki (2007). (In the graph $f(R)$ denotes our $f(R) - R$, and f_{R0} is our $f'(R) - 1$ today.) The ratio c_1/c_2 is fixed such that $f(R)$ approaches for large R/m^2 the Λ CDM value. The value of R/m^2 at the present epoch is about 40. Thus only the part of the graph to the right of this value is relevant.



Since this model will also later play a role, we present it here. Its analytic form reads

$$f(R) = R - m^2 \frac{c_1(R/m^2)^n}{c_2(R/m^2)^n + 1}, \quad m^2 := \frac{\kappa^2 \rho_0}{3} \tag{10.16}$$

(see Figure 10.1), with suitably chosen parameters c_1, c_2 (n is a positive integer, and ρ_0 denotes the present average cosmic density). The corresponding history leads to a curvature equation of state parameter $w_{\text{eff}}(z)$ that deviates from -1 , but these deviations can be kept sufficiently small. At high redshifts $w_{\text{eff}}(z)$ becomes smaller than -1 , a possibility that can be checked with future observations. Such a crossing of the so-called phantom line would be interesting. We do not discuss here the evolution of linear cosmological perturbations for $f(R)$ models that may imply interesting testable deviations from the concordance model. (For some references on this, see the last section.)

10.4 Weak-Field Limit for Spherically Symmetric Sources

We remarked earlier that for $f(R)$ gravity the Schwarzschild or Schwarzschild-de Sitter metric is often a vacuum solution, e.g., for the model (10.16). This does, however, not guarantee that the theory passes solar system tests. This would be the case if this vacuum solution can be matched to an interior solution (as Schwarzschild showed for GR). We shall see that this is generically not possible. In this section we address this issue in the weak-field limit.

As a preparation we linearize the scalar field equation (10.6) about a background de Sitter universe with $\Lambda = R_s/4$, where the Ricci scalar R_s satisfies

$$f'_s R_s - 2f_s = 0 \tag{10.17}$$

($f_s := f(R_s)$, etc). Let $R(r) = R_s + \delta R(r)$, and linearize the trace equation for a local source:

$$3f_s''\square\delta R + (R_s f_s'' - f_s')\delta R = \kappa^2 T. \quad (10.18)$$

Since T vanishes for the background, we regard it as of first order. The last equation shows that the scalar field δR has an effective mass given by⁷

$$m_s^2 = \frac{f_s' - f_s'' R_s}{3f_s''} = \frac{(f_s')^2 - 2f_s f_s''}{3f_s' f_s''}. \quad (10.19)$$

After considerable confusion, it was shown in Chiba et al. (2007) that the post-Newtonian Eddington-Robertson parameter γ is not equal to 1, as in GR and also observationally to high accuracy, but $\gamma = 1/2$, if the following conditions are satisfied:

- (i) Linearization of $f(R)$, $f'(R)$ about R_s is allowed.
- (ii) $f''(R_s) \neq 0$.
- (iii) The Compton wave length $1/m_s$ is much larger than the size of the solar system.
- (iv) The deviations of the gravitational field from the de Sitter background metric can be treated in first order.

Remarks These conditions are not always satisfied. If $f''(R_s) = 0$, then $\gamma = 1$ as in GR. Condition (iii) can be violated, for instance, by fine-tuning the parameters in

$$f(R) = R + \frac{1}{\alpha^2} R^2 - \mu^4/R.$$

The only way to escape the destructive consequence $\gamma = 1/2$ and maintain the late cosmic acceleration is to invoke a ‘‘chameleon mechanism’’ for the scalar degree of freedom.

10.5 Chameleon Mechanism

The chameleon effect was discovered by Khoury and Weltman (2004a,b) in scalar field models of DE. Scalar fields with self-interactions may directly couple to matter as strong (or even stronger) as gravity and still satisfy all current constraints. The reason for this is that the effective mass of the scalar field depends on the local density. So there is the possibility that the Compton wavelength is sufficiently small

⁷It turns out (Faraoni 2005) that the nonnegativity of the expression in (10.19) is the stability condition of the de Sitter spacetime with respect to small inhomogeneous perturbations of the $f(R)$ model (without matter).

on Earth to satisfy all laboratory bounds, while it is much larger in the solar system and still much larger on cosmological scales.

For illustration, consider a scalar field model satisfying the nonlinear equation

$$\square\varphi = V'_{\text{eff}}(\varphi), \quad V_{\text{eff}} = V(\varphi) - B(\beta\varphi)T, \quad (10.20)$$

where T is the trace of the matter part of the energy-momentum tensor ($\approx -\rho$ if the pressure can be neglected). The dependence of V_{eff} on ρ can imply that $\partial^2 V_{\text{eff}}/\partial\varphi^2$ at the effective minimum is much smaller for a low-density background than in a high-density environment (see Figure 10.2). This density dependence can lead to a *thin-shell effect*: φ varies for a macroscopic body only over a thin surface layer, leading to a weak fifth force. This behavior is intimately related to the nonlinear nature of Chameleon field theories. An equation of the type (10.20) is obtained for $f(R)$ models in the Einstein frame. By transforming Equation (10.7), one finds

$$\tilde{\square}\varphi = \frac{dU}{d\varphi} + \frac{1}{2}\beta e^{-2\beta\varphi}T.$$

Only nonrelativistic matter contributes to T . If we define $\hat{\rho}$ by

$$T \approx -\rho =: -e^{(3\beta/2)\varphi}\hat{\rho},$$

then $\hat{\rho}$ is conserved in the Einstein frame. In terms of this quantity, we obtain

$$\tilde{\square}\varphi = \frac{dU_{\text{eff}}}{d\varphi}, \quad U_{\text{eff}}(\varphi) = U(\varphi) + e^{-(\beta/2)\varphi}\hat{\rho},$$

which is of the form (10.20) with an exponential function B (as in Figure 10.2). In what follows, we do not make use of this Einstein frame formulation.

The possibility of a Chameleon mechanism for the scalar degree of freedom of $f(R)$ gravity models has been studied in several papers, e.g., in Navarro and Van Acoleyen (2007) and Faulkner et al. (2007). We discuss here briefly part of the work by Hu and Sawicki (2007).⁸

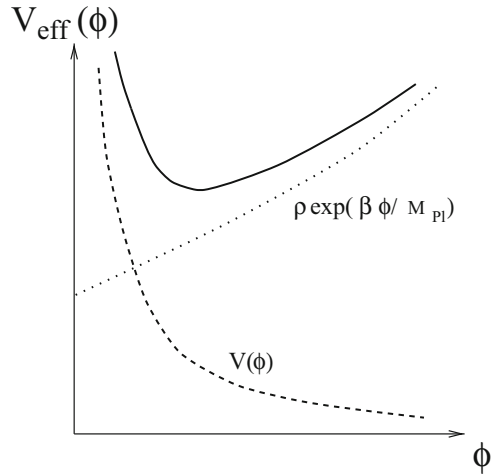
These authors study in the framework of $f(R)$ models nonrelativistic stars like the Sun with weak gravitational fields, but do no more linearize the equation for $\phi = f'(R)$. For a static situation, Equation (10.6) reduces to

$$3\nabla^2 f'(R) + Rf'(R) - 2f(R) = \kappa^2 T \approx -\kappa^2 \rho. \quad (10.21)$$

It is a good approximation to replace the Laplacian of the slightly curved spatial metric by the flat space Laplacian (for which we use the same symbol). Given a density profile $\rho(r)$ from a solar model, Equation (10.21) becomes a nonlinear

⁸For a simplified discussion in the Einstein frame, see Capozziello and Tsujikawa (2008).

Fig. 10.2 Typical effective potential of the form (10.20), whose density dependence can lead to a Chameleon effect. From Brax et al. (2004).



field equation for R (or ϕ). Hu and Sawicki choose the model (10.16) and impose the following boundary conditions: Deep inside the star f' assumes the value with $R = \kappa^2 \rho$ (implied by GR). Very far away ($r = 10^6 r_\odot$), the outer boundary condition $f' = f'(R = \kappa^2 \rho_g)$ is imposed, where ρ_g is the average galactic density in the solar vicinity ($\rho_g = 10^{-24} \text{ g cm}^{-3}$). These boundary conditions correspond approximately to the minima of the effective potential belonging to (10.21).⁹ At this point we consider them as part of the model. The chosen density profile is shown in Figure 10.3 (solid line), while the numerical solution of the boundary value problem for $R(r)$ is shown by the dashed line. (The parameter f_{R0} in this figure is $f' - 1$ for the present average cosmic scalar Ricci curvature.) A blown-up version of the region where R does not track the GR limit $\kappa^2 \rho$ outside about 1 AU is shown in Figure 10.4.

Once $\rho(r)$ and $R(r)$ are known, the field equations (10.3) determine the Einstein tensor, from which the metric in the weak-field limit can easily be computed (Poisson integrals).

Since R deviates from the GR value $\kappa^2 \rho$ only in a very low-density shell, the solar system tests present no problem. For example, the γ parameter becomes

$$\gamma \approx 1 - \frac{2M_{\text{eff}}}{3M + M_{\text{eff}}}, \tag{10.22}$$

where M is the total mass of the star, and

$$M_{\text{eff}} = 4\pi \int (\rho - R/\kappa^2)r^2 dr. \tag{10.23}$$

⁹The effective potential is defined by $\frac{\partial V_{\text{eff}}}{\partial \phi} = -\frac{\kappa^2}{3}\rho + \frac{1}{3}(2f - Rf')$.

Fig. 10.3 Realistic density profile of the solar interior and vicinity (solid curve), and solution R/κ^2 of the scalar field equation for $n = 4$ and field amplitude $|f'(R_0) - 1| = 0.1$, with boundary conditions described in the text (dashed line). From Hu and Sawicki (2007), with the same change of notation as in Figure 10.1.

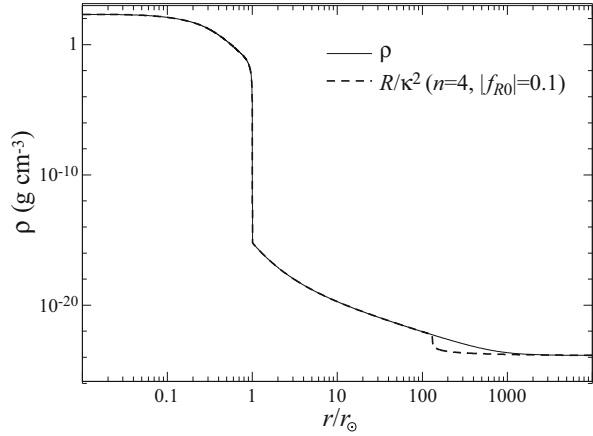
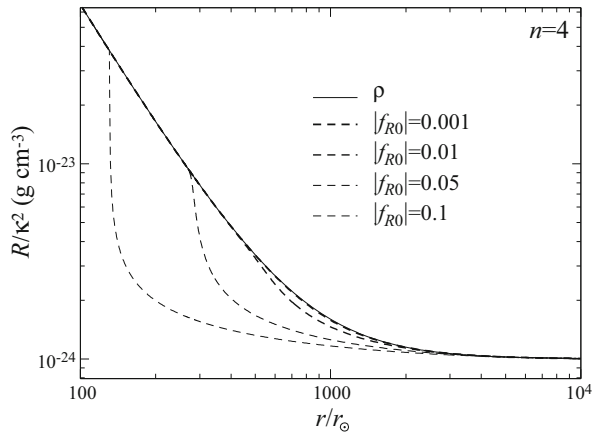


Fig. 10.4 Blown-up version of Figure 10.3 for various choices of the cosmological field amplitude $f'(R_0) - 1$ (equal to f_{R0} in the notation of Hu and Sawicki 2007).



For the solutions shown in Figures 10.3 and 10.4 $M_{\text{eff}} \ll M$, thus $\gamma \approx 1$ to high accuracy. So this looks very good, but it has to be emphasized, that this conclusion rests on the assumption, that the galaxy is in the high-curvature regime ($R \approx \kappa^2 \rho$) with respect to its own density profile. The validity of this assumption depends, as Hu and Sawicki stress, “on both the structure of the galactic halo and its evolution during the acceleration epoch.” This dependence is irritating, but reflects the fact that there is *no Birkhoff theorem* for $f(R)$ gravity models. To decide about the proper boundary conditions, one would have to study – horrible dictu – galaxy formation for $f(R)$ models in N-body simulations.

We shall see in the next section that the Chameleon mechanism can only work if the star has a surrounding medium, e.g., dark matter.

10.6 Nonexistence of Relativistic Stars in $f(R)$ Gravity?

Based on the nonlinear scalar field equation (6), Frolov recently gave analytic arguments that for strong-field matter configurations, curvature singularities may quite generally develop (Frolov 2008). This suggestion has been investigated by Kobayashi and Maeda (2008) in a numerical study of stellar models for a family of $f(R)$ theories which incorporate the Chameleon mechanism. The result of this study is that the considered family cannot describe relativistic stars. It remains to be seen how general this important result is, but it suggests that $f(R)$ gravity as a viable modifications of GR may well be excluded. In this section we briefly describe the content of the paper by Kobayashi and Maeda (abbreviated as KM).

Especially for the numerical part of their work, KM use the following model:

$$f(R) = R + \lambda \bar{R} \left[\left(1 + \frac{R^2}{\bar{R}^2} \right)^{-n} - 1 \right] \quad (\lambda, \bar{R} > 0), \quad (10.24)$$

that was adopted by Starobinsky (2007) to incorporate the Chameleon mechanism. Since only the qualitative behavior of the potential¹⁰ $V(\phi) := \mathcal{U}/\phi^2$ near the de Sitter minimum matters, the results will also apply to (16) and other models (Appleby & Battye 2007). We set the de Sitter value $R_s = x_s m$, then λ is uniquely determined by x_s (for a given integer n). The same holds for the de Sitter value ϕ_s of ϕ . We note that

$$\phi(R) = 1 - 2n\lambda \frac{R}{\bar{R}} \left(1 + \frac{R^2}{\bar{R}^2} \right)^{-n-1}, \quad (10.25)$$

which shows that a curvature singularity ($R \rightarrow \pm\infty$) is mapped to $\phi = 1$.

It should be remarked that $V(\phi)$ is a multivalued function, but this is no worry because only a relatively small interval about ϕ_s really matters.

For this model KM study spherically symmetric stars. The metric is parameterized in Schwarzschild coordinates as

$$g = -N(r)dt^2 + \frac{1}{B(r)}dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2). \quad (10.26)$$

It is easy to generalize the GR structure equations to nonlinear $f(R)$ models. The (tt) and (rr) components of the field equations (10.5) become

$$\frac{\phi}{r^2}(-1 + B + rB) = -8\pi GV - \phi^2 V - B \left[\phi_{rr} + \left(\frac{2}{r} + \frac{B_r}{2B} \right) \phi_r \right], \quad (10.27)$$

$$\frac{\phi}{r^2}(-1 + B + rB \frac{N_r}{N}) = 8\pi GP - \phi^2 V - B \left(\frac{2}{r} + \frac{N_r}{2N} \right) \phi_r. \quad (10.28)$$

¹⁰Note that $V(\phi)$ is closely related to U in (10.11). It is easy to see that the de Sitter value for R is mapped to the value of ϕ , where V takes its minimum.

An index r denotes differentiation with respect to r . The scalar field equation (10.7) becomes

$$B \left[\phi_{rr} + \left(\frac{2}{r} + \frac{N_r}{2N} + \frac{B_r}{2B} \right) \phi_r \right] = \frac{8\pi G}{3} (-\rho + 3P) + \frac{2\phi^3}{3} V'(\phi). \quad (10.29)$$

Recall that the energy-momentum “conservation” gives, as in GR,

$$P_r + \frac{N_r}{2N} (\rho + P) = 0. \quad (10.30)$$

We turn to the boundary conditions. Near the center of the star at $r = 0$, the various functions are expanded in powers of r^2 , making also use of the scaling freedom of the time coordinate. Some of the coefficients are determined in terms of others by the field equations. So far KM have assumed, for simplicity, that the energy density of the star is constant, thus generalizing Schwarzschild’s interior solution. Given ρ and the central values P_c, ϕ_c of P and ϕ , the basic Equations (10.27)–(10.29) can be integrated outwards from the center to the surface $r = R_*$ of the star, which is defined by $P(R_*) = 0$. Note that Equation (10.30) gives

$$N(r) = \left[\frac{\rho + P_c}{\rho + P(r)} \right]^2. \quad (10.31)$$

From the surface the vacuum equations are integrate to sufficiently large r . The starting values at the center are then varied until the de Sitter solution with $\phi \rightarrow \phi_s$, as $r \rightarrow \infty$ is assumed.

It turns out that there are no solutions if the gravitational potential $\Phi := (1 - N)/2$ is larger than some value Φ_{max} , which is typically about 0.1. (Recall that for GR $\Phi_{max} = 4/9$.) When KM tried to find solutions with larger Φ , the *Ricci scalar diverged*. Thus, for the studied class of $f(R)$ models, there are *no stars with strong gravitational fields*. The authors give also analytic arguments for this conclusion that were originally suggested by Frolov (2008). These are based on a mechanical interpretation of the scalar field equation (10.29).

For nonrelativistic stars, it turns out that the thin-shell condition is violated, when there is no matter outside the stellar surface, and therefore the parameter γ is close to 0.5. For such stars one can also derive good analytic approximations.

In the previous section, we saw that the Chameleon effect can work for nonrelativistic stars if surrounding matter is taken into account. Surrounding matter, however, does not change the nonexistence statement for strong gravitational fields, as is shown by KM. In view of these results $f(R)$, gravity models are in serious trouble. More realistic equations of state will presumably not change this conclusion.

10.7 Inclusion of Other Curvature Invariants

There are a number of studies (e.g., Carroll et al. 2005), that include other curvature invariants, such as $R_{\mu\nu}R^{\mu\nu}$, $R_{\alpha\beta\gamma\delta}R^{\alpha\beta\gamma\delta}$. Such models are in most cases *unstable*, like mechanical Lagrangian systems with higher derivatives (Woodard 2007).¹¹ An exception seems to be Lagrangians which are functions of R and the Gauss-Bonnet invariant $R_{GB} = R^2 - 4R_{\mu\nu}R^{\mu\nu} + R_{\alpha\beta\gamma\delta}R^{\alpha\beta\gamma\delta}$. By introducing two scalar fields, such models can be written as an Einstein-Hilbert term plus a particular extra piece, containing a linear coupling to R_{GB} . Because the Gauss-Bonnet invariant is a total divergence, the corresponding field equations are of second order. This does, however, not guarantee that the theory is ghost-free. In De Felice et al. (2006) this question was studied for a class of models (Carroll et al. 2005) for which there exist accelerating late-time power-law attractors. It turned out that in a Friedmann background, there are no ghosts, but there is instead *superluminal propagation* for a wide range of parameter space. This acausality is reminiscent of the Velo-Zwanziger phenomenon (Velo and Zwanziger 1969a,b) for higher (> 1) spin fields coupled to external fields. It may very well be that it can only be avoided if very special conditions are satisfied. Ghosts of Gauss-Bonnet cosmologies have also been studied in Calcagni et al. (2006).

In addition to these problems, it appears unlikely that the devastating difficulties we have encountered for $f(R)$ models will disappear when other curvature invariants are included.

10.8 First-Order (Affine) Modifications of GR

The disadvantage of complicated fourth-order equations can be avoided by using the *Palatini variational principle*, in which the metric and the symmetric affine connection ($\Gamma^\alpha_{\mu\nu}$) are considered to be independent fields.¹²

For GR the “Palatini formulation” is equivalent to the Einstein-Hilbert variational principle, because the variational equation with respect to $\Gamma^\alpha_{\mu\nu}$ implies that the affine connection has to be the Levi-Civita connection. Things are no more that simple for $f(R)$ models:

$$S = \int \left[\frac{1}{2\kappa^2} f(R) + L_{matter} \right] \sqrt{-g} d^4x, \quad (10.32)$$

¹¹This paper contains a discussion of a generic instability of Lagrangian systems in mechanics with higher derivatives that was discovered by Ostrogradski (1850).

¹²This approach was actually first introduced by Einstein (1925). This is correctly stated in Pauli’s classical text on relativity (p. 215).

where $R[g, \Gamma] = g^{\alpha\beta} R_{\alpha\beta}[\Gamma]$, $R_{\alpha\beta}[\Gamma]$ being the Ricci tensor of the independent torsionless connection Γ . The equations of motion are in obvious notation:

$$f'(R)R_{(\mu\nu)}[\Gamma] - \frac{1}{2}f(R)g_{\mu\nu} = \kappa^2 T_{\mu\nu}, \quad (10.33)$$

$$\nabla_{\alpha}^{\Gamma} (\sqrt{-g} f'(R) g^{\mu\nu}) = 0. \quad (10.34)$$

For the second of these equations, one has to assume that L_{matter} is functionally independent of Γ . (It may, however, contain metric covariant derivatives.)

Equation (10.34) implies that

$$\nabla_{\alpha}^{\Gamma} \left[\sqrt{-\hat{g}} \hat{g}^{\mu\nu} \right] = 0 \quad (10.35)$$

for the conformally equivalent metric $\hat{g}_{\mu\nu} = f'(R)g_{\mu\nu}$. Hence, the $\Gamma^{\alpha}_{\mu\nu}$ are equal to the Christoffel symbols for the metric $\hat{g}_{\mu\nu}$.

The trace of (10.33) gives

$$Rf'(R) - 2f(R) = \kappa^2 T.$$

Thanks to this algebraic equation, we may regard R as a function of T : $R = \mathcal{R}(T)$. In the matter-free case, it is identically satisfied if $f(R)$ is proportional to R^2 . Otherwise R is equal to a constant c (which is in general not unique). If $f'(c) \neq 0$, Equation (10.34) implies that Γ is the Levi-Civita connection of $g_{\mu\nu}$, and (10.33) reduces to Einstein's vacuum equation with a cosmological constant. In general, one can rewrite the field equations in the form of Einstein gravity with nonstandard matter couplings.¹³

$$\begin{aligned} f'G_{\mu\nu}[g] &= \kappa^2 T_{\mu\nu} - \frac{1}{2}(\mathcal{R}f' - f)g_{\mu\nu} \\ &- \frac{3}{2f'}[\nabla_{\mu} f' \nabla_{\nu} f' - \frac{1}{2}g_{\mu\nu}(\nabla f')^2] + (\nabla_{\mu} \nabla_{\nu} - g_{\mu\nu} \square) f'. \end{aligned} \quad (10.36)$$

With this reformulations it is, for instance, straightforward to develop cosmological perturbation theory (Koivisto and Kurki-Suonio 2006).

For some time this modification of GR looked promising. But now one can ignore it because of the following major drawbacks.

1. Since the vacuum theory is identical with GR including a cosmological constant, the metric g outside a spherically symmetric star has to be the Schwarzschild-de Sitter metric. This does, however, not guarantee that the solar system tests are

¹³It is shown in Sotiriou (2006) that if the matter action is independent of Γ , the theory is dynamically equivalent to a Brans-Dicke theory with the special Brans-Dicke parameter $-3/2$, plus a potential term.

satisfied. For this we have to know whether there are interior solutions that match the exterior metric field. This was studied in several papers, with a negative result for physically relevant equations of state. Technically it was shown in Barausse et al. (2008a,b) that for polytropic equations of state with an adiabatic index Γ in the interval $3/2 < \Gamma < 2$ true curvature singularities develop. From this one can guess with confidence that the *theory cannot describe white dwarfs* – for example.

2. For nonlinear Palatini $f(R)$ gravity the *Cauchy problem is not well-posed* (Lanahan-Tremblay and Faraoni 2007).¹⁴

Both of these unacceptable shortcomings have in the final analysis a common origin. The good thing about the field equations in the form (10.36) is that they are of second order in the derivatives of the metric. What leads to the mentioned difficulties is that the right-hand side of (10.36) is at least of second order in the matter variables, because R is a function of T . Apart from ideal fluids, T is usually of first order, and then the right-hand side is even of third order in the matter fields. This induces locally sensitive dependence of the metric on rapidly changing matter fields, in contrast to GR (and Newtonian gravity) where such dependencies are smoothed out.

10.9 Concluding Remarks

A positive aspect of the largely negative outcome of the previous discussion seems to me that the distinguished role of GR among large classes of gravity theories has once more become apparent. We know, of course, that GR is an effective theory, and that quantum theory will produce all sorts of induced terms (a phenomenon that is well-known from QED), but stopping any expansion after a few terms will hardly lead to a consistent theory that agrees with observations on all scales.

Some of the modified gravity theories, such as $f(R)$ or braneworld models, may perhaps be of limited use for testing GR on cosmological scales. Guided by such models,¹⁵ there have recently been some interesting attempts to develop a parameterized post-Friedmann description of gravity that parallels the parameterized post-Newtonian description of solar system tests (discussed by C. Will at this meeting). In contrast to the latter, there appear unavoidably some free functions, instead of just a bunch of parameters, in the description of the evolution of inhomogeneities (Hu 2008; Bertschinger & Zukin 2008). It will, therefore, be difficult to discriminate between dark energy and modified gravity, but this remains

¹⁴In Lanahan-Tremblay and Faraoni (2007), it is shown that the basic system of equations in vacuum cannot be rewritten as a system of only first order, since $\square\phi$ cannot be eliminated, except of course if $\square\phi = 0$ (e.g., for the vacuum theory).

¹⁵Especially from the evolution of linear cosmological perturbations for such models (Song et al. 2006; Lanahan-Tremblay and Faraoni 2007; Bean et al. 2007; Pogosian and Silvestri 2007).

a major goal for years to come. One can hope that this will eventually become possible with better data on the CMB background, weak gravitational lensing, and the growth of large-scale structures.

References

- Amendola, L., Gannouji, R., Polarski, D., & Tsujikawa, S. (2007). Conditions for the cosmological viability of $f(R)$ dark energy models. *Physical Review D*, 75, 083504. arXiv:gr-qc/0612180.
- Appleby, S. A., & Battye, R. A. (2007). Do consistent $F(R)$ models mimic general relativity plus Λ ? *Physical Letters B*, 654, 7. arXiv:0705.3199 [astro-ph].
- Barausse, E., Sotiriou, T. P., & Miller, J. C. (2008a). A no-go theorem for polytropic spheres in Palatini $f(R)$ gravity. *Classical Quantum Gravity*, 25, 062001. arXiv:gr-qc/0703132.
- Barausse, E., Sotiriou, T. P., & Miller, J. C. (2008b). Curvature singularities, tidal forces and viability of Palatini $f(R)$ gravity. *Classical Quantum Gravity*, 25, 105008. arXiv:0712.1141.
- Bean, R., Bernat, D., Pogosian, L., Silvestri, A., & Trodden, M. (2007). Dynamics of linear perturbations in $f(R)$ gravity. *Physical Review D*, 75, 064020. arXiv:astro-ph/0611321.
- Bertschinger, E., & Zukin, P. (2008). Distinguishing modified gravity from dark energy. *Physical Review D*, 78, 024015. arXiv:0801.2431 [astro-ph].
- Brax, Ph., van de Bruck, C. Davis, A.-C., Khoury, J., & Weltman, A. (2004). Chameleon dark energy. *Physical Review D*, 70, 123518. arXiv:astro-ph/04101103.
- Calcagni, G., de Carlos, B., & De Felice, A. (2006). Ghost conditions for Gauss-Bonnet cosmologies. *Nuclear Physics B*, 752, 404. arXiv:hep-th/0604201.
- Capozziello, S., & Tsujikawa, S. (2008). Solar system and equivalence principle constraints on $f(R)$ gravity by chameleon approach. *Physical Review D*, 77, 107501. arXiv:0712.2268 [gr-qc].
- Caroll, S. M., De Felice, A., Duvvuri, V., Easson, D., Trodden, M., & Turner, M. S. (2005). Cosmology of generalized modified gravity models. *Physical Review D*, 70, 063513. arXiv:astro-ph/0410031.
- Chiba, T., Smith, T. L., & Erickcek, A. L. (2007). Solar system constraints to general $f(R)$ gravity. *Physical Review D*, 75, 124014. arXiv:astro-ph/0611867.
- Copeland, E. J., Sami, M., & Tsujikawa S. (2006). Dynamics of dark energy. *International Journal of Modern Physics D*, 15, 1753. arXiv:hep-th/0603057.
- De Felice, A., Hindmarsh, M., & Trodden, M. (2006). Ghosts, instabilities, and superluminal propagation in modified gravity models. *Journal of Cosmology and Astroparticle Physics*, 0608, 005. arXiv:astro-ph/0604154.
- Einstein, A. (1925). Einheitliche Feldtheorie von Gravitation und Elektrizität. In *S. B. Preuss. Akad. Wiss.* (pp. 414–419). Hoboken: Wiley
- Faraoni, V. (2005). Modified gravity and the stability of de Sitter space. *Physical Review D*, 72, 061501(R). arXiv: gr-qc/0509008.
- Faulkner, T., Tegmark, M., Bunn, E. F., & Mao, Y. Constraining $f(R)$ gravity as a scalar tensor theory. *Physical Review D*, 76, 063505. arXiv:astro-ph/0612569.
- Fay, S., Nesseris, R., & Perivolaropoulos, L. (2007). Can $f(R)$ gravity theories mimic a Λ CDM cosmology? *Physical Review D*, 76, 063504. arXiv:gr-qc/0703006.
- Frolov, A. V. (2008). A singularity problem with $f(R)$ dark energy. *Physical Review Letters*, 101, 061103. arXiv:0803.2500 [astro-ph].
- Hu, W. (2008). Parameterized post-Friedmann signatures of acceleration in the CMB. *Physical Review D*, 77, 103524. arXiv:0801.2433 [astro-ph].
- Hu W., & Sawicki, I. (2007). Models of $f(R)$ cosmic acceleration that evade solar-system tests. *Physical Review D*, 76, 064004. arXiv:0705.1158 [astro-ph].

- Khoury, J., & Weltman, A. (2004a). Chameleon fields: Awaiting surprises for tests of gravity in space. *Physical Review Letters*, *93*, 171104. arXiv:astro-ph/0309300.
- Khoury, J., & Weltman, A. (2004b). Chameleon cosmology. *Physical Review D*, *69*, 044026. arXiv:astro-ph/0309411.
- Kobayashi, T., & Maeda, K. (2008). Relativistic stars in $f(R)$ gravity, and absence thereof. *Physical Review D*, *78*, 0644019. arXiv:0807.2503.
- Koivisto T., & Kurki-Suonio, H. (2006). Cosmological perturbations in the Palatini formulation of modified gravity. *Classical Quantum Gravity*, *23*, 2355. arXiv:astro-ph/0509422.
- Lanahan-Tremblay, N., & Faraoni, V. (2007). The Cauchy problem of $f(R)$ gravity. *Classical Quantum Gravity*, *24*, 5667. arXiv:0709.4414.
- Maeda, K. i. (1989). Towards the Einstein-Hilbert action via conformal transformation. *Physical Review D*, *39*, 3159.
- Navarro, I., & Van Acoleyen, K. (2007). $f(R)$ actions, cosmic acceleration and local tests of gravity. *Journal of Cosmology and Astroparticle Physics*, *0702(022)* (2007). arXiv:gr-qc/0611127.
- Ostrogradski, M. (1850). *Memoire Academie St. Petersburg, Ser. VI* (Vol. 4, pp. 385).
- Pauli, W. (1985–99). In K. von Meyenn (Ed.). *Wissenschaftlicher Briefwechsel mit Bohr, Einstein, Heisenberg u.a.* (Vol. 1–4). New York: Springer.
- Pogosian L., & Silvestri, A. (2007). The pattern of growth in viable $f(R)$ cosmologies. *Physical Review D*, *77*, 023503. arXiv:0709.0296.
- Salgado, M. (2006). *Classical Quantum Gravity*, *23*, 4719 (2006).
- Salgado, M., Rio, D. M.-d. Alcubierre, M., & Nunez D. (2008). *Physical Review D*, *77*, 104010. arXiv:0801.2372 [gr-qc].
- Song, Y., Sawicki, I., & Hu, W. (2006). The large scale structure of $f(R)$ gravity. *Physical Review D*, *75*, 064003. arXiv:astro-ph/0606286.
- Sotiriou, T. P. (2006). $f(R)$ gravity and scalar-tensor theory. *Classical Quantum Gravity*, *23*, 5117. arXiv:gr-qc/0604028.
- Sotiriou T. P., & Faraoni V. (2010). $f(R)$ theories of gravity. *Reviews of Modern Physics*, *82*, 451–497. arXiv:0805.1726 [gr-qc].
- Starobinsky, A. A. (2007). Disappearing cosmological constant in $f(R)$ gravity. *Journal of Experimental and Theoretical Physics Letters*, *86*, 157. arXiv:0706.2041 [astro-ph].
- Straumann, N. (2007). Dark energy. In *Approaches to fundamental physics*. In E. Seiler & I.-O. Stamatescu (Eds.), *Lecture Notes in Physics* (Vol. 721, pp. 327–397). Berlin: Springer.
- Velo, G., & Zwanziger, D. (1969a). Propagation and quantization of Rarita-Schwinger waves in an external electromagnetic potential. *Physical Review*, *186*, 1337–1341.
- Velo, G., & Zwanziger, D. (1969b). Noncausality and other defects of interaction Lagrangians for particles with spin one and higher. *Physical Review*, *188*, 2218–2222.
- Whitt, B. (1984). Fourth order gravity as general relativity plus matter. *Physics Letters B*, *145*, 176.
- Woodard, R. P. (2007). Avoiding dark energy with $1/R$ modifications of gravity. *Lecture Notes in Physics*, *720*, 403. arXiv:astro-ph/0601672.

Chapter 11

The Unexpected Resurgence of Weyl Geometry in late 20th-Century Physics



Erhard Scholz

11.1 Introduction

In the 1970s three groups of authors, essentially independent of one another, began to reconsider Weyl's generalization of Riemannian geometry from 1918. Weyl had proposed the latter from the perspective of building a geometrically unified theory of gravitation and electromagnetism. By the end of the 1920s, after the successful reformulation of the underlying gauge idea in relativistic quantum physics, most physicists including Weyl himself had given up on the idea of extending the geometry of spacetime by means of a "localized" scaling degree of freedom. It was hardly to be expected that half a century later researchers of the next generation would try again to give Weyl geometry a new role within the changed context of late 20th century physics. But some of them did. A first group of authors, in particular J. Ehlers, F. Pirani, and A. Schild, used it as a conceptual framework for clarifying the foundations of gravity; another group explored extended gravitational theories using the generalized geometrical structure, and still another, including W. Drechsler and H. Tann in Munich, investigated connections between gravity and quantum physics. In several of these approaches a scalar field extending the gravitational structure played a crucial role. Although none of these attempts found an immediate broader response, some of them led to follow up papers. As a result, different research perspectives exploring questions of recent physics from a Weyl-geometric viewpoint emerged, though they remained too heterogeneous to coalesce into a coherent literary tradition or to form a common research community.

The call for papers for the Mainz conference proceedings provided a splendid incentive for taking stock of the broader range of Weyl-geometric investigations

E. Scholz (✉)

University of Wuppertal, Department C, Mathematics and Natural Sciences, and Interdisciplinary Center for Science and Technology Studies, Wuppertal, Germany

e-mail: scholz@math.uni-wuppertal.de

in physics that emerged during the last three decades of the 20th century. Of course the following survey cannot be complete; rather it has to be confined within specified boundaries. So this paper is restricted to the more classical parts of gravity together with some relatively limited attention given to connections with quantum theory. *Not covered* in this survey is the whole range of Weyl-geometric methods in Kaluza-Klein theories, supergravity, and in string theory.

In order to facilitate the reading of the following survey, the paper starts with a very short introduction to, or a reminder of, central features of Weyl geometry and gravity (section 1.1). Because a considerable portion of the developments following utilize a scale-covariant scalar field coupled to the Hilbert term, similar to the one in Jordan–Brans–Dicke (JBD) gravity, the second part of the first section is devoted to a short glance at JBD theory from a Weyl–geometric perspective (section 1.2). The other sections give a partly historical, partly systematic survey of attempts to use Weyl-geometric methods in recent physics.

In section 2 three different, partially overlapping, approaches from the 1970s are described. The already mentioned paper of Ehlers, Pirani and Schild (EPS) on the foundations of gravity and some follow up papers are dealt with in section 2.1. A completely different take arose from proposals put forward by a group of Japanese physicists, M. Omote, R. Utiyama et al. and independently by P.A.M. Dirac. They investigated a scalar field coupling to the Hilbert term similar to JBD gravity, but in the scale-invariant approach of Weyl geometry. The interpretation of the Weylian scale connection by Dirac was not the same as that of the Japanese physicists; they and their immediate successors had different research contexts in mind. In Dirac's case, the context was gravity, astrophysics, cosmology and electromagnetism; in Utiyama's, it was nuclear and elementary particle physics (section 2.2). Finally, although less noticed in the wider community, a specific road to Weyl-geometric structures arose in research on gauge theories of gravity arising from the Kibble-Sciama program for deriving gravitational structures (fields) from “localizing” symmetries in Minkowski space, often considered from a wider perspective than that of the Poincaré group. In this view, Weyl geometry arose as a special case of Cartan geometry, since Weyl-geometric gravity ought to be generically extended by a translational connection component, viz. torsion. It is a surprising fact that these three revivals of Weyl-geometric gravity, although arising from completely different backgrounds and pursuing different goals, were undertaken and published in the short period from 1971 to 1974, exactly when the basics for the standard model of elementary particle physics were established (section 2.3).

Before coming to the follow up investigations which made use of these approaches in the standard model of elementary particle physics and/or in astrophysics and cosmology, we turn towards an even more surprising recourse to Weyl geometry in attempts to geometrize quantum mechanics (QM) in the wake of the Bohmian heterodoxy (section 3). In order to make this kind of geometrization accessible to readers not versed in Bohmian quantum mechanics, the basic ideas necessary to understand the geometrization proposals are briefly reviewed in section 3.1. This is followed by a survey of the offbeat road leading to a geometrization of configurations spaces in QM by Weyl geometry, as developed

in the 1980s by E. Santamato's and continued after 2010 with his colleague F. De Martini (section 3.2). A more delicate idea of a Bohm-type quantization procedure in cosmology leading to a Weyl-geometric framework, as proposed by A. and F. Shohai and M. Golshani, is the topic of section 3.3.

In section 4 we turn towards different attempts at using Weyl-geometric structures (mainly scale invariance and the scale-invariant affine connection) and fields (Weylian scale connection and/or an additional scale-covariant scalar field) in elementary particle physics. Two interrelated questions arise naturally if one wants to bring gravity closer to the physics of the standard model (SM):

- (i) Is it possible to bring conformal, or at least scale-covariant generalizations of classical (Einsteinian) relativity into a coherent common frame with the standard model SM?¹
- (ii) Is it possible to embed classical relativity in a quantized theory of gravity or, the other way round, to derive classical relativity as an effective theory arising from a more fundamental quantum gravity theory at the classical level?

The fact that all the SM fields, with the sole exception of the Higgs field, have conformally invariant Lagrangians in the context of special relativity, i.e., Minkowski space, was considered already in the mid-1970s by F. Englert and coauthors, among others. This circumstance seemed to cry out for investigations from a Weyl-geometric perspective which then, of course, would invite generalizing the spacetime environment of all SM fields, at least in their pseudo-classical form,² to Lorentzian or Weylian manifolds. In this context the Weylian scale connection was identified by L. Smolin at the end of the 1970s as a new hypothetical field, which after quantization would lead to a particle with mass close to the Planck scale. Roughly ten years later this hypothetical particle was independently rederived by H. Cheng and called a "Weylon" (section 4.1). Roughly a decade later the question of "mass generation" by breaking the scale symmetry in a Weyl-geometric approach to SM fields was again studied in Munich by W. Drechsler and H. Tann (section 4.2).

This question continued to attract the interest of researchers at least until the empirical detection of the Higgs boson in 2012. In the last few years in particular, H. Nishino and S. Rajpoot, but not only they, have studied the question of how the symmetry of the standard model may be enhanced by a scale degree of freedom and may be broken by a peculiar interplay between an initially scale-covariant scalar field and the "Weylon". All this was discussed at the pseudo-classical level

¹Such an attempt seemed to be supported experimentally by the phenomenon of (Bjorken) scaling in deep inelastic electron-proton scattering experiments. The latter indicated, at first glance, an active scaling symmetry of mass/energy in high energy physics; but it turned out to hold only approximatively and was of restricted range.

²SM fields are here called *pseudo-classical* if they are considered before, or better abstracting from (so-called "second") quantization. Mathematically they are classical fields (spinor fields or gauge connections), but the field components do not correspond to physically measurable quantities. Observationally relevant information can be extracted only after applying perturbative quantization methods or, in semiclassical approximations, from their wave-function-like probability currents.

(section 4.3). In the recent years, some authors have turned towards difficult questions that arise from Weyl scaling at the quantum level. A group of Italian authors, G. Codello, G. D’Orico, C Pagani, and R. Percacci brought forward new arguments with regard to the commonly shared view that scale symmetry seems to be broken at the quantum level. They have proposed procedures under which scale invariance can be preserved under quantization. H. Ohanian has recently discussed the transition from a scale-invariant phase of fields close to the Planck scale to a lower energy regime with broken scale symmetry and Einstein gravity as effective field theory (section 4.4).

The rescaling allowed in Weyl geometry may change the geometrical picture underlying our usually assumed cosmological models. Scalar fields with conformal rescaling have been in use for a long time in “early universe” modelling (section 5.1). They invite Weyl-geometric investigations, an approach dominated for several decades by N. Rosen and M. Israelit, the former an early protagonist of the Dirac approach to Weyl-geometric gravity. In the last few decades other authors have jumped in with slightly different ideas (section 5.2f). A coherent tradition with a larger group of researchers in astrophysical and cosmological studies has formed in Brazil around M. Novello. They invoke a (weak) Weyl-geometric framework and have been pursuing its implications for more than two decades, This Brazilian group has had the largest stable set of contributors, more than any other line of research considered in this survey (section 5.4). But the question of dark matter effects, if considered from the gravitational side, has to be counted among the successes of modified Newtonian dynamics, MOND, and this aspect has remained outside the scope of the Brazilian research program. First steps at reconstructing MOND-like phenomenology in a Weyl-geometric approach to gravity, made recently at Wuppertal, seem sufficiently striking to be included here (section 5.3).

A survey of a side-stream in recent physical research, as attempted in this paper, cannot claim to tell a coherent story or even point to clear successes. Its goal is rather to collect views from necessarily heterogeneous perspectives in order to bring them together in a single panorama. In this way it invites the reader to look backward and forward and thereby reflect on the development of methods and views in recent mathematical and theoretical physics (section 6).

11.2 Preliminaries: Weyl-Geometric Gravity and Jordan-Brans-Dicke Theory

11.2.1 Weyl Geometry and Gravity

11.2.1.1 Basics of the Geometrical Framework

Weyl geometry is a generalization of Riemannian geometry, arising from two insights: (i) The mathematical automorphisms of Euclidean geometry and of special relativity are the *similarities* (of Euclidean, or respectively of Lorentz

signature) rather than the congruences. No unit of length is naturally given in Euclidean geometry, and likewise the basic structures of special relativity (inertial motion and causal structure) can be given without the use of clocks and rods.³ (ii) The development of field theory and general relativity demands a conceptual implementation of this insight in a strictly *localized mode* (physics terminology).⁴ In more physical language, (i) and (ii) can be given the form of a postulate stating that fundamental field theories have to be formulated covariantly under point dependent rescalings of the basic units of measurement, while the Lagrangian densities and the dynamical laws (the “natural laws”) remain invariant under these rescalings (see Dicke’s postulate cited in Section 11.2.2). It remains an open question whether the resulting extension of the mathematical automorphism groups for such theories will be of physical import, or whether it amounts to a purely mathematical refinement.

Based on these insights, Weyl developed what he called purely infinitesimal geometry (*reine Infinitesimalgeometrie*) building upon a conformal generalization of a (pseudo-) Riemannian metric g with coefficient matrix $(g_{\mu\nu})$ and with (point-dependent) rescaling $\tilde{g}(x) = \Omega(x)^2 g(x)$ (Ω a nowhere vanishing strictly positive function), along with a scale (“length”) connection given by a real-valued differential form $\varphi = \varphi_\mu dx^\mu$ (Weyl 1918a,c). If one rescales the metric by Ω , one has to *gauge transform* φ by $\tilde{\varphi} = \varphi - d \log \Omega$. The *scale connection* (φ_μ) expresses a comparison of lengths of vectors (or other metrical quantities) at two infinitesimally close points, both measured in terms of a representative $(g_{\mu\nu})$ of the conformal class. The typical symmetry of the geometry at the infinitesimal level is thus the scale-extended Poincaré group, sometimes called the *Weyl group* (although the same name is used in Lie group theory in a completely different sense). From 1918 to roughly 1921/22 it seemed clear to Weyl that this extension of Riemannian geometry could be used for unifying gravity and electromagnetism; later he gave up this hope and considered his scale geometry as a purely mathematical enterprise, the most important features of which were transplanted to the $U(1)$ -gauge theory of electromagnetism.⁵

With hindsight, Weyl’s generalization of Riemannian geometry may be seen as embedded in E. Cartan’s even wider program of geometries with infinitesimal symmetries. In the case of the scale-extended Poincaré group one then arrives at a Cartan-Weyl geometry with a translational Cartan connection and *torsion* as the

³This distinguishes Euclidean geometry from the other classical geometries of constant curvature. The consideration of material systems, like hydrogen or caesium atoms etc., may be used to introduce units and reduce the allowed automorphisms of the congruences.

⁴In mathematical terminology, the implementation of a similarity structure happens at the *infinitesimal* level. In the following the physicists’ use of the terminology “local” will be used. A discussion, given by Weyl later in his life, of the role of mathematical and physical automorphisms can be found in (Weyl 1949/2016), some aspects of this also appear in (Weyl 1949, chap. III, sec. 14).

⁵For more historical and philosophical details see, among others, (Vizgin 1994; Goenner 2004; Ryckman 2005; Scholz 1999; Scholz et al. 2001).

typical extension of the structure.⁶ With the exception of Section 11.3.3 this paper will be restricted to the original form of Weyl geometry without torsion; large parts of it, in fact, only deal with the simplest case, that of an integrable scale connection. The reasons for this restriction will become apparent below.

Metrical quantities in Weyl geometry are directly comparable only if they are measured at the same point p of the manifold. Quantities measured at different points $p \neq q$ that are a non-infinitesimal distance apart can only be compared metrically after an integration of the scale connection along a path from p to q . Weyl realized that this structure is compatible with a uniquely determined affine connection $\Gamma = (\Gamma_{\nu\lambda}^\mu)$, the *affine connection of Weyl geometry*. If the Levi-Civita connection of the Riemannian part g is denoted by ${}_g\Gamma_{\nu\lambda}^\mu$, the Weylian affine connection is given by

$$\Gamma_{\nu\lambda}^\mu = {}_g\Gamma_{\nu\lambda}^\mu + \delta_\nu^\mu \varphi_\lambda + \delta_\lambda^\mu \varphi_\nu - g_{\nu\lambda} \varphi^\mu. \quad (11.1)$$

In the following, the *covariant derivative* with respect to Γ will be denoted as $\nabla = \nabla_\Gamma$. Similarly, the curvature expressions for the Riemann tensor, Ricci tensor and scalar curvature $Riem$, Ric , R will denote the corresponding Weylian entities. The corresponding scale gauge dependent Riemannian analogues derived from ${}_g\Gamma_{\nu\lambda}^\mu$ will be written as ${}_g\nabla$, ${}_gRiem$, ${}_gRic$, ${}_gR$. The Weylian scalar curvature, e.g., is

$$R = {}_gR - (n-1)(n-2) \varphi_\mu \varphi^\mu - 2(n-1) {}_g\nabla_\mu \varphi^\mu, \quad (11.2)$$

where n is the dimension of the manifold. A change of scale neither changes the connection (the left hand side of (11.1)) nor the covariant derivative; only the composition from the underlying Riemannian part and the corresponding scale connection (right hand side) is shifted.

As every connection defines a unique curvature tensor, curvature concepts from “ordinary” (Riemannian) differential geometry pass over to Weyl geometry. The Riemann and Ricci tensors, $Riem$, Ric , are scale invariant by construction, although their expressions contain terms in φ . On the other hand, the scalar curvature involves a “lifting” of indices by the inverse metric and is thus scale covariant of weight -2 (see below). For vector and tensor fields (i.e. quantities that are not dimensionless) the appropriate scaling behaviour under change of the metrical scale has to be taken into account. If a field, expressed by X (leaving out indices) with respect to the metrical scale $g(x)$, transforms like $\tilde{X} = \Omega^k X$ with regard to the scale choice $\tilde{g}(x)$ as above, then X is called a *scale-covariant* field of *scale weight*, or *Weyl weight* $w(X) := k$ (usually an integer or a fraction). This is the negative value of the mass weights used in particle physics. In general the covariant derivative, ∇X , of a scale-covariant quantity X will not be scale covariant; nevertheless, scale covariance can

⁶For a modern presentation of Cartan geometry, including the Cartan-Weyl case, see, e.g., (Sharpe 1997, chap. 7); for the physical aspects of the extension studied since the 1970s (Blagojević/Hehl 2013, chap. 8).

be recovered by adding a weight-dependent term. The *scale-covariant derivative* D of X is defined by $DX := \nabla X + w(X)\varphi \otimes X$, or using a coordinate description

$$D_\mu X^\nu := \nabla_\mu X^\nu + w(X)\varphi_\mu X^\nu. \quad (11.3)$$

For example, the derivative ∇g is not scale covariant, but Dg is – in fact, it even vanishes:

$$Dg = \nabla g + 2\varphi \otimes g = 0 \quad (11.4)$$

In Weyl geometry the metric is thus no longer *constant* with respect to the derivative ∇ , but it is *with respect to the scale-covariant derivative* D . From the point of view of Riemannian geometry this appears as a “non-metricity” of the connection (in the literature often called “semi-metricity”), whereas seen from the perspective of Weyl geometry this is nothing but the *metric compatibility condition* for Γ .

This should suffice for recalling the fundamental properties. More details on Weyl geometry can be found in Weyl’s original papers (Weyl 1918a,c), those of his younger contemporaries (Eddington 1923; Bergmann 1942; Schouten 1924, 1954; Dirac 1973) and in the more recent literature.⁷

11.2.1.2 Weyl-Geometric Gravity

Weyl’s original generalization of Riemannian geometry arose from an attempt to reformulate Einstein’s theory of gravity so as to obtain a geometrical unification of gravity and electromagnetism (Vizgin 1994; Goenner 2004). Any meaningful Lagrangian in this framework will be subject to the constraint of scale symmetry. Since $w(\sqrt{|g|}) = 4$, whereas the Weyl-geometric scalar curvature R has weight $w(R) = -2$, Weyl could not work with the Hilbert-Einstein term. He therefore considered quadratic expressions in the curvature terms to obtain an appropriate generalization of the gravitational Lagrangian, e.g.⁸

$$\mathcal{L}_{\mathfrak{W}} = L_W \sqrt{|g|} \quad \text{with } L_W = \alpha_1 R^\mu_{\nu\lambda\kappa} R^\nu{}^\lambda{}_\mu{}^\kappa + \alpha_2 R^2. \quad (11.5)$$

⁷Presentations of Weyl geometry can be found, among others, in (Blagojević 2002; Israelit 1999b), (Tonnelat 1965, chap. IX), (Drechsler/Tann 1999, appendix A) and (Perlick 1989) (difficult to access). For selected aspects see (Codello et al. 2013) and (Ohanian 2016, sec. 4). Integrable Weyl geometry is presented in (Dahia et al. 2008; Romero et al. 2011; Almeida et al. 2014; Quiros 2014b), (Scholz 2011a, sec. 2.1). Be aware of different conventions for the scale connection. Expressions for Weyl-geometric derivatives and curvature quantities are derived in (Gilkey et al. 2011; Yuan/Huang 2013) and (Miritzis 2004, App.). For a more mathematical perspective consult (Folland 1970; Calderbank/Pedersen 1998; Gauduchon 1995; Higa 1993; Ornea 2001; Gilkey et al. 2011).

⁸As the “most simple and natural” expression $\alpha_2 = 0$ in (Weyl 1918a) and $\alpha_1 = 0$ as the simplest example in (Weyl 1918b, 4th ed., 5th ed.).

Of course, he added a term for the scale curvature $f = d\varphi$ ($f_{\mu\nu} = \partial_\mu\varphi_\nu - \partial_\nu\varphi_\mu$), which looks like the Maxwell action:

$$\mathfrak{L}_f = L_f \sqrt{|g|} \quad \text{with } L_f = -\frac{1}{4} f_{\mu\nu} f^{\mu\nu} \quad (11.6)$$

Only much later – in fact, about half a century later – did theorists begin to consider other gravitational Lagrangians with Weyl-geometric scale symmetry. These arose from the idea of a coupling between gravity, here the Weyl-geometric scalar curvature R , and a scalar field with the “correct” complementary weight (Section 11.3.2).

In the period covered here we encounter two different modes of Weyl-geometric gravity. One, sometimes called *Weyl gravity* (in the strong sense), is farther removed from Einstein gravity and uses square curvature Lagrangians. The other, closer to Einstein gravity, works with a modified Hilbert term coupled to a scalar field; in the physics literature it is often called *Weyl-geometric scalar-tensor theory* (WST). The latter goes back to independent proposals by M. Omote and R. Utiyama, on the one hand, and P.A.M. Dirac, on the other, that try to make use of a scalar field modification of the Hilbert term, analogous to Jordan-Brans-Dicke theory (see Section 11.3.2). Here the gravitational structure is characterized by an equivalence class of triples (g, φ, ϕ) , where $g = g_{\mu\nu} dx^\mu dx^\nu$ is the *Riemannian component* of the Weylian metric, $\varphi = \varphi_\mu dx^\mu$ its scale connection, and ϕ an additional scalar field.⁹ The equivalence is given by combined rescaling transformations: $g \mapsto \tilde{g} = \Omega^2 g = e^{2\omega} g$, $\varphi \mapsto \tilde{\varphi} = \varphi - d\omega$, $\phi \mapsto \tilde{\phi} = e^{-\omega} \phi$. Because of scaling freedom, a Weylian metric with nowhere vanishing scalar curvature can be gauged to $R \doteq \text{const}$ (here, and elsewhere in this paper, \doteq denotes an equality which holds only in a certain gauge specified by the context). Weyl considered this as the “natural” gauge, but we prefer to call it the *Weyl gauge*.

If the scale connection is an exact form,

$$\varphi = -dw, \quad (11.7)$$

with a scalar potential w scale transforming by $w \mapsto \tilde{w} = w + \omega$, then this is an *integrable Weyl-geometric scalar-tensor theory* (IWST). In such a case, the gravitational structure reduces to the Riemannian component of the metric plus, at face value, two scalar fields $(g, \phi = e^v, w)$ with equivalence under rescaling. As $v \mapsto \tilde{v} = v - \omega$, the sum $v + w$ is a *scale-invariant* scalar field of the gravitational structure, in fact the *only crucial* one. Because of the scale gauge freedom ϕ , v or w can, respectively, be given any chosen value, e.g., a constant.¹⁰ In the integrable case, two scale gauges are of particular importance in addition to the Weyl gauge: the *Riemann gauge*, in which the scale connection is “integrated away” (for $\omega = -w$),

⁹In a way, this may be called a geometrical “tensor-vector scalar” theory *sui generis*, in which all components have geometrical meaning.

¹⁰The dynamical consequences of this interdependence have been clarified by (Israelit 1999b,a), see Section 11.6.2.1.

then $\tilde{\varphi} \doteq 0$. The other is the *scalar-field gauge* in which the scalar field is scaled to a constant, $\tilde{\varphi}(x) \doteq \phi_o = \text{const}$ (for $\omega = v$). If the value of ϕ_o is specified so that it hooks up with Einstein gravity, $\phi_o = (8\pi G)^{-1}$ (up to a hierarchy factor if need be), it is called the *Einstein gauge*.¹¹

For a vanishing scale-invariant sum,

$$v + w = 0, \quad (11.8)$$

the scalar field ϕ is essentially the potential for the scale connection, more precisely,

$$\varphi = dv = d \ln \phi. \quad (11.9)$$

It is then and only then that the Einstein and Riemann gauge will coincide and IWST reduces to Einstein gravity. The Palatini approach, varying the metric and the affine connection of a Lagrange density $\mathcal{L} = \phi^2 R \sqrt{|g|}$ independently, enforces both the constraint $v + w = 0$ and the integrability of the scale connection. This implies that a Palatini-IWST will reduce to Einstein gravity. The latter is then only re-written in scale-covariant form, but without any modification of the dynamics.¹² If one considers IWST from the point of view of the metric-affine scheme, then it is better to use variational constraints as in (Cotsakis/Miritzis 1999) rather than the Palatini approach. This does not force the condition (11.8), and the scalar field, respectively the integrable scale connection (11.9), will then express an additional dynamical degree of freedom.

11.2.2 Jordan-Brans-Dicke (JBD) Gravity

11.2.2.1 Basics of JBD Theory

In the early 1950s *Pascual Jordan* (Hamburg) and, a decade later, *Carl Brans* and *Robert Dicke* (Princeton) proposed a generalization of Einstein gravity by considering a variable gravitational parameter.¹³ Their motivations were essentially different, but there was nevertheless a strong overlap that led to the ensuing theory, here abbreviated by JBD. Jordan started from an action principle (Jordan 1952, p. 140)

$$\mathcal{L}_g(\chi, g) = (\chi R - \frac{\xi}{\chi} \partial^\mu \chi \partial_\mu \chi) \sqrt{|det g|}, \quad (11.10)$$

¹¹Obviously the Einstein gauge exists also in the non-integrable case.

¹²Cf. Sections 11.6.2.2, 11.6.4.

¹³A similar approach had been proposed by W. Scherrer in the 1930s but had received no attention by other scientists (Goenner 2012). For recent surveys on scalar tensor theories including JBD gravity see (Capozziello/Faraoni 2011, chap. 3), (Clifton/Ferreira et al. 2012, chap. 3.1)

with a parameter ξ and a real scalar field χ functioning as a kind of spacetime-dependent (reciprocal) gravitational “constant”, and with R the usual Riemannian scalar curvature of the metric g (Jordan 1952, 2nd. ed., 163, (3)).¹⁴ A Lagrange term L_m for classical matter was foreseen in connection with this formalism (e.g., (Brans 1961, eqn. (6))). The hypothesis of a “varying gravitational constant” had been conjectured already more than a decade earlier by P.A.M. Dirac as part of his speculations on “large numbers” and their relations in physics.¹⁵ Pauli reminded Jordan that his “extended gravity” allowed for a class of conformal transformations, which not only affect the metric but also the scalar field,

$$\tilde{g}_{\mu\nu} = \Omega^2 g_{\mu\nu}, \quad \tilde{\chi} = \Omega^{-2} \chi. \quad (11.11)$$

Jordan included this generalization in the second edition of his book (Jordan 1952, 2nd ed., 169).¹⁶

A few years later, Carl Brans (Brans 1961) and Robert Dicke took up the study of scale-covariant scalar fields (including a classical matter term L_m in (11.10) (Brans 1961, 8)). Their motivation was to formulate a theory of gravity which took account of Mach’s principle as understood by D.W. Sciama.¹⁷ For the two US physicists the main function of the scalar field was “the determination of the local value of the gravitational constant” (Brans 1961, 929). More clearly than in Jordan’s work, they emphasized the wave character of the dynamical equation of χ (Brans 1961, eqs. (9), (13)). Moreover, they had a different view of the role of scale transformations: their methodological goal was to establish a scale-independent foundation for physical theories.

This involved a “passive” interpretation of scale transformations (Brans 1961, 927), whereas Jordan and Pauli tended to think in terms of “active” scale transformations of material structures. Dicke began an article dedicated to *transformations of units* in GRT (Dicke 1962) by announcing the following principle:

It is evident that the particular values of the units of mass, length, and time employed are arbitrary and that the laws of physics must be invariant under a general coordinate-dependent transformation of units. (Dicke 1962, 2163)

This was very much in the spirit of Weyl’s original intentions in 1918, but not with his views afterward, that is once he had disassociated himself from his

¹⁴Warning: One has to check carefully the sign convention used in the definition of Riemann and scalar curvature. Jordan, e.g., used sign inverted definitions of the curvature terms with respect to those used here and in much of the present literature (Jordan 1952, 40). In Fujii/Maeda’s notation (see below) this would correspond to $\epsilon = -1$ and thus to a “ghost” field.

¹⁵For Dirac’s role in this story see (Kragh 2016), and for a larger view of JBD theory see (Brans 1999, 2014).

¹⁶The conformal factor Ω was (unnecessarily) restricted by the condition $\Omega^2 = \chi^\gamma$ for some constant $\gamma \in \mathbf{R}$.

¹⁷In the 1950s Sciama had considered the possibility that the gravitational “constant” was related to the mass and the “radius” of the visible universe.

gravitational program. After his gauge idea had shifted over to quantum physics, Weyl discussed this new viewpoint on different occasions in the 1940s, e.g. in (Weyl 1949/2016, p. 165).¹⁸ It seems that Brans and Dicke “reinvented” the idea of scale gauge invariance of the natural laws anew. They systematically discussed the scale transformations of physical quantities, based on the (quasi-axiomatic) principle of the invariance of the velocity of light c and the Planck constant \hbar . In particular, he noted that “all three quantities, time, length, and reciprocal mass transform in the same way” (Dicke 1962, 2164), i.e.,

$$l' = \Omega l, \quad t' = \Omega t, \quad m' = \Omega^{-1} m.$$

In this sense, Weyl’s scale gauge transformations reappeared in the context of Jordan-Brans-Dicke theory without being explicitly mentioned as such. It may well be that at the time no one but Pauli was aware of this close resemblance to Weyl’s theory. In effect, Weyl’s choice of (scale) gauge was translated by Dicke into the choice of a *frame* of measuring units, complementing the choice of a coordinate system.

In more recent papers the scalar field and the JBD parameter are usually written in a slightly different form. Taking $\phi = \sqrt{2\xi^{-1}\chi}$, scale weight $w(\phi) = -1$, and $\xi = \frac{\epsilon}{4\omega}$, the Lagrangian (11.10) turns into

$$\mathcal{L}_{BD} = \left(\frac{1}{2}\xi\phi^2 R - \frac{1}{2}\epsilon\partial^\mu\phi\partial_\mu\phi + L_{mat} \right) \sqrt{|det g|}, \quad (11.12)$$

where $\text{sig } g = (3, 1) \sim (-+++)$ and $\epsilon = 1$, except for exceptional cases where $\epsilon = -1$ or 0 (Fujii/Maeda 2003, p. 5).¹⁹ In the following discussion this notation will be used as standard.

A famous exception in which $\epsilon = -1$ is the special constellation of coefficients

$$L_{cc} = \phi^2 R + 6\partial_\mu\phi\partial^\mu\phi. \quad (11.13)$$

¹⁸Also in the English edition of *Philosophy of Mathematics and Natural Sciences* Weyl expressed this disassociation quite clearly, appealing to the constants of atomic physics which regulate the frequencies of spectral lines (Weyl 1949, 83). But this was only one part of his perspective. In the appendix he argued that for a deeper insight it would be necessary to understand how the “adaptation” of the mass of the electron to the local field constellation is achieved (Weyl 1949, 288f). This was close to the intentions of his 1918 approach, although no longer a claim that the goal had been achieved. Einstein, in his later papers, agreed (Einstein 1949, 555f); see (Lehmkuhl 2014).

¹⁹ $\epsilon = 1$ corresponds to a normal field having a positive energy, in other words, not a “ghost”. Fujii/Maeda add that $\epsilon = -1$ looks unacceptable because it seems to indicate negative energy, but “this need not be an immediate difficulty owing to the presence of the nonminimal coupling” (ibid.).

In this case, the Lagrangian is invariant (up to an exact differential) under conformal transformations (Penrose 1965), which produces *conformal coupling*. This framework can be used to study gravitational theories “in which scale invariance of matter is a consistency requirement on its coupling to gravitation” (Deser 1970). Deser considered a conformally coupled scalar field as a paradigmatic example for matter, and he furthermore observed that the addition of a quadratic term of the form $\frac{1}{2}\mu^2\phi^2$, with μ a parameter of mass dimension, implies *breaking* of the *conformal symmetry*. In this case, the range ϕ “must clearly be cosmological in order not to lead to a clash with observation” (Deser 1970, p. 252).

This contrasts with the view of the three founding authors – Jordan, Brans and Dicke – who considered it as evident that the “conformal transformations” (scale transformations) do *not* reduce the geometrical considerations to those of a purely conformal structure. For them it was clear that JBD theory possesses a covariant derivative ∇ , specified by the reference metric g underlying (11.10), which was fundamental for Jordan as well as Brans/Dicke. Later this scale was called the *Jordan frame* (although Jordan was undecided as to which scale might be the “natural”one). Because the Levi-Civita connection of the Jordan frame metric determines the free fall trajectories of test particles, many authors consider this to be the “physical frame”, whereas the other frames are then only mathematical auxiliary devices. On the other hand, the JBD-field ϕ can be scaled to a constant. If this is done, then the gravitational part of the Lagrangian looks like the Hilbert term of Einstein gravity, while the remnants of the JBD scalar field appear in additional expressions of the Lagrange density.²⁰ The resulting *Einstein frame* will then satisfy the Riemannian “energy conservation” condition for matter tensors. Since roughly the 1990s this approach has found an increasing number of adherents who take this to be the proper frame for a “physical” interpretation of JBD gravity. Still, no consensus has emerged in the JBD community, so this question has remained undecided, to say the least (Faraoni/Nadeau 2007; Quiros et al. 2013).

11.2.2.2 JBD in a Weyl-Geometric Perspective

The perspective of (integrable) Weyl geometry may help clarify certain aspects underlying this debate. Let us denote the affine connection referred to by

$$\nabla := {}_g\nabla, \quad (11.14)$$

where the r.h.s. expresses the Levi-Civita connection of g in the JBD Lagrangian (11.10). ∇ is kept unaffected, i.e. *invariant*, under scale transformations in JBD theory. A structural view of Weyl geometry shows that the combination of a conformal structure $[g]$ of pseudo-Riemannian metrics g and a specification of an invariant affine connection ∇ with a compatibility condition, built in here because of

²⁰See, e.g., (Capozziello/Faraoni 2011, chap. 3.6).

(11.14), determines a Weyl structure on a differentiable manifold M .²¹ In this way JBD gravity may be embedded in the theoretical frame of Weyl geometry. Most authors of JBD theory are not aware of this structural feature of their theory. In the last few years, however, at least two authors have hinted at such a connection (see section 5.2.2).²²

11.3 Contributions to Weyl-Geometric Gravity in the 1970s and 1980s

11.3.1 Ehlers/Pirani/Schild and Subsequent Work

11.3.1.1 An Axiomatic Approach to the Foundations of Gravity

Already in the fourth edition of *Raum – Zeit – Materie* Weyl discussed the relationship between the physical concept of a *causal structure* and the mathematical concept of a *conformal structure* on a differentiable manifold (Weyl 1918b, 4th. ed., appendix I). He deemed it inadequate to think of a direct empirical determination of the metrical coefficients $g_{\mu\nu}$ by “rods and clocks,” so he looked for another empirical specification of a Weylian metric (g, φ) . In a note added to a letter to F. Klein²³ (a little later published in *Göttinger Nachrichten* as (Weyl 1921)) he sketched an idea showing how this could be achieved. Starting from the framework of “purely infinitesimal” geometry, Weyl showed that if two generalized metrics have identical conformal structures and induce the same projective geodesic path structures then they will coincide. This meant that the conformal and projective path structures suffice to determine a Weylian metric uniquely.²⁴

Fifty years later, Weyl’s argument combining projective and conformal structures was taken up again and extended by Jürgen Ehlers, Felix Pirani and Alfred Schild (EPS in the sequel) (Ehlers et al. 1972). This was around the same time that Dirac was using Weyl geometry in the context of scalar-tensor theories. The EPS paper was written for a *Festschrift* honouring J.L. Synge, who was known among other things for proposing that general relativity be based on the behaviour of standard clocks (*chronometric approach*). From the foundational point of view, clocks might appear to be a problematic choice considering that they involve complicated material systems. The question thus arose whether more basic signal structures from gravitational theory (light rays, particle trajectories) might be adequate. In the words of Hilbert, EPS “laid the foundations deeper”, combining Weyl’s ideas from 1920/21

²¹ See, e.g., (Gilkey et al. 2011).

²² See the discussion in (Quiros et al. 2013) and (Scholz 2017).

²³ (Weyl 1920)

²⁴ Weyl’s note (Weyl 1921) became better known by his calculation and discussion of projective and conformal curvature tensors, which followed.

with the recently developed mathematical language and symbolic technology of differentiable manifolds, but doing so in the spirit of Hilbert's axiomatic method.²⁵

They started from three sets, $\mathcal{M} = \{p, q, \dots\}$, $\mathcal{L} = \{L, N, \dots\}$, $\mathcal{P} = \{P, Q, \dots\}$, with $\mathcal{L}, \mathcal{P} \subset \mathcal{M}$, calling these collections of *events*, *light rays* and *particles*. Next came their properties, formulated as postulates based on experimental findings for light signal exchange between particles. These were set down in different groups of axioms ($D_1, \dots, D_4, L_1, L_2, P_1, P_2, C$), very much in the Hilbertian style of foundations of geometry. This enabled them to introduce a C^3 differentiable structure on \mathcal{M} for which \mathcal{L} and \mathcal{P} described smooth curves (axiom group D). Moreover, \mathcal{L} (axioms L) defined a C^2 conformal structure, while \mathcal{P} (axioms P) induced a differentiable projective path structure. The compatibility axiom C essentially postulated that light rays can always be approximated arbitrarily well by particle trajectories. Putting these conclusions together, led to the main result.

Theorem 1 (Ehlers/Pirani/Schild 1972) *A light ray structure \mathcal{L} and a set of particle trajectories \mathcal{P} defined on an event set \mathcal{M} which satisfy axioms D, L, P, C endow \mathcal{M} with the structure of a (C^3 -) differentiable manifold M and a (C^2 -) Weylian metric $[(g, \varphi)]$. The latter is uniquely determined by the condition that its causal and geodesic structures coincide with \mathcal{L} and \mathcal{P} respectively.*

EPS also posed the question: how might the (pseudo-)Riemannian structure of classical (Einsteinian) relativity arise from this Weylian one? A simple additional *Riemannian axiom*, postulating the vanishing of the scale curvature, $d\varphi = 0$, would serve that purpose. Such a postulate did not seem implausible, especially since Weyl's interpretation of the scale connection φ as electromagnetic was by now obsolete and EPS did not adhere to it. Nevertheless, these authors did not wish to exclude the possibility that a scale connection field φ with nonvanishing scale curvature might play the role of a "true" field, even if still unknown.

11.3.1.2 Subsequent Work

The paper of Ehlers, Pirani and Schild triggered a line of investigations in the foundations of general relativity. This is sometimes called the *causal inertial approach* (Coleman/Korté), but sometimes it is subsumed under the more general search for a *constructive axiomatics* of GRT (Majer/Schmidt, Audretsch, Lämmerzahl, Perlick and others). These investigations turned towards a basic conceptual analysis from the point of view of foundations of inertial geometry (Coleman/Korté 1984), some even exploring a Desarguesian type characterization of free fall lines (Pfister 2004). In (Perlick 1989, 1987, 1991) an attempt was made to introduce "standard clocks" in the Weyl-geometric setting without taking refuge in atomic processes but rather by using observations of light rays and inertial trajectories. An extremely

²⁵See (Trautman 2012).

well-considered discussion of the measurability question in relation to the EPS analysis of the conformal and free fall structure of spacetime has recently been given by O. Darrigol (Darrigol 2014, chap. 5.2/5.3).²⁶

Another line of follow up work explored an extension of the causal inertial approach to quantum physics, where particle trajectories no longer seemed acceptable as a foundational concept. This led to a debate, opened by Jürgen Audretsch, Konstanz, (Audretsch 1983) and immediately continued by the collective work of three authors (Audretsch/Gähler/Straumann 1984), cited in the sequel by AGS, and in a series of follow up studies.²⁷ Audretsch argued that the gap between Weylian and Riemannian geometry can “be closed if quantum theory as a theory of matter is made part of the total scheme” (Audretsch 1983, 2872). He postulated that quantum theory, in the sense of Dirac or Klein-Gordon (K-G) fields on a Weylian manifold, is compatible with a Weyl geometry if and only if the WKB (Wentzel-Kramers-Brillouin) approximation of the Dirac (or K-G) field leads to streamlines, which in the limit $\hbar \rightarrow 0$ agree with the geodesics of a Riemannian metric (Audretsch’s *compatibility condition*).

In the follow up work, cited in footnote 27, Audretsch and collaborators hoped to strengthen their arguments by referring to recent neutron and atomic interference experiments in the gravitational field of the earth and/or in accelerated frames of reference. They put forward their own version of a “constructive axiomatics” for spacetime based on the analysis of the classical limit of the wave mechanics of matter fields. They even exploited the possibility of using a Cartan-type extension of the underlying Riemannian geometry for modelling the propagation of the spin field associated with a matter wave (Audretsch/Lämmerzahl 1988, 1994). In collaboration with F. Hehl (Cologne) they argued that for the propagation of the spin current $S^\mu = \bar{\psi} \gamma_5 \gamma^\mu \psi$ of a Dirac field ψ , considered as a pseudo-classical field, an (axial) vector field on spacetime played a role, which they described as a torsion component added to the Levi-Civita connection of the Riemannian structure.²⁸

Working with scale-covariant mass factors m of Weyl weight $w(m) = -1$, Audretsch found that compatibility is possible only if the mass factor m of the Dirac particle leads to a vanishing covariant derivative $\nabla_\mu m = 0$ in some gauge. He observed that this implied the vanishing of Weyl’s scale curvature $d\varphi = 0$ (Audretsch 1983, eqn. (6.14)) and concluded a bit rashly:

The consequence of the requirement is therefore that the Weyl space reduces to a Riemann space and the gap [between Weylian and Riemannian geometry, ES] described in Sec. I is closed. (Audretsch 1983, 2881, emph. in original)

²⁶I thank S. Walter for the hint.

²⁷Among them in particular (Audretsch/Lämmerzahl 1988; Lämmerzahl 1990; Audretsch/Lämmerzahl 1991; Audretsch/Hehl/Lämmerzahl 1992) and (Audretsch/Lämmerzahl 1994).

²⁸(Audretsch/Hehl/Lämmerzahl 1992, pp. 390ff), compare (Audretsch/Lämmerzahl 1988, eqn. (4.5)).

In fact, Audretsch had only shown that the limiting condition for streamlines of the WKB approximation for the Dirac field to classical geodesic trajectories implied *integrability* of the Weylian metric, since the vanishing of $\nabla_\mu m$ in the Riemann gauge implies that $D_\mu m = 0$ in any gauge. The question as to whether this meant that the Riemann gauge should be regarded as “physical” was not posed; this was rather imputed as self-evident.

In the AGS paper this question was taken up again, but stated more carefully in the language of conformal fibre bundles for Dirac- and for Klein-Gordon fields. AGS showed that Audretsch’s compatibility condition implies the possibility of reducing the “conformal” group, here understood as $\mathbf{R}^+ \times SO(1, 3)$, to the orthogonal group. Their central argument was that the WKB (Wentzel-Kramers-Brillouin) approximation of a (locally defined) Dirac or Klein-Gordon field ψ on M on a Weylian manifold leads to flow-lines which, in the limit $\hbar \rightarrow 0$, agree with the Weylian geodesics if and only if the Weyl metric is integrable.²⁹

All in all, the three authors formulated their results with more care than Audretsch had done in his first paper. They did not claim that their investigation had completely filled the gap between Weylian and Riemannian geometry, although it had reduced its depth considerably. It might seem natural to choose the Riemann gauge of the Weyl metric in order to reduce the structure group to $SO(3, 1)$, but nothing compels us to do so. The classical interpretation of geodesics as trajectories of mass points was, in any case, foreign to the field theoretic context. In place of this, one now postulated coherence between geodesic structure and the flow-lines associated with pseudo-classical quantum fields.

11.3.2 Dirac’s and Omote/Utiyama’s Retake of Weyl Geometry

11.3.2.1 Dirac on Scale-Covariant “Varying” Gravity

In the 1970s *P.A.M. Dirac* introduced Weyl geometry into the discourse of the rising scalar-tensor theories. He was still driven by his fascination with certain constellations of large numbers in physics, the “large number hypothesis” (Dirac 1973, 1974).³⁰ Mainly following Eddington’s notation and terminology (Eddington 1923), Dirac introduced readers from the younger generation of physicists to Weyl

²⁹The flow could be characterized by the 0-th order probability current j_o^μ of the pseudo-classical field. In their argument that the current is geodesic the authors neglected terms proportional to j_o^μ (Audretsch/Gähler/Straumann 1984, eqn.(5.14), (5.16)). The repercussions of this generosity remains unclear to me (ES).

³⁰Dirac presented his proposal for a retake of Weyl geometry at the occasion of the symposium honouring his 70th birthday, 1972 at Trieste. This talk remained unpublished. According to (Charap/Tait 1974, p. 249 footnote) the talk was close to his 1973 publication. For the broader historical context of this enterprise, the background in Dirac’s reflection on large numbers in the 1920s, and a surprising link to geophysics see (Kragh 2016).

geometry, since this theory was no longer generally known. He then developed a scalar-tensor theory of gravitation coupled in an “old-fashioned”, i.e. outdated, way to electromagnetism. Just as Weyl did in 1918, Dirac now identified the electromagnetic field $F_{\mu\nu}$ with the Weylian scale curvature $f = d\varphi$

$$F_{\mu\nu} = f_{\mu\nu} .$$

Henceforth, I will refer to this as the *electromagnetic (em) dogma*. On the other hand, Dirac replaced Weyl’s original Ansatz for gravity in the Lagrangian (using square curvature terms) by a JBD-type Lagrangian with a real scalar field β of weight $w(\beta) = -1$. He also added a quartic scale-invariant potential term,

$$\mathcal{L}_{Dir\ 0} = -\beta^2 R + k D^\lambda \beta D_\lambda \beta + c\beta^4 + \frac{1}{4} f_{\mu\nu} f^{\mu\nu} , \quad (11.15)$$

with constant k . For $k = 6$, the scale connection terms of the Lagrangian essentially cancel (i.e., they reduce to boundary terms and thus are variationally negligible) like in (11.13). So Dirac wrote the Lagrangian in the form

$$\mathcal{L}_{Dir\ 1} = -\beta^2 {}_gR + 6\partial^\lambda \beta \partial_\lambda \beta + c\beta^4 + \frac{1}{4} f_{\mu\nu} f^{\mu\nu} , \quad (11.16)$$

which is known to be conformally invariant (Penrose 1965); here ${}_gR$ denotes the *sign inverted* Riemannian scalar curvature with respect to the generally accepted convention, while in (11.15) R is the sign inverted scalar curvature of the Weylian metric.³¹

Dirac derived dynamical equations and Noether identities for diffeomorphisms and scale transformations. He distinguished the Riemann gauge (called “natural gauge” by him), $\varphi = 0$, which arises, of course, only for a vanishing e.m. field $f_{\mu\nu} = 0$ from the “Einstein gauge” (gravitational parameter constant, $\beta = 1$) and the “atomic gauge” based on atomic clocks. He warned that “all three gauges are liable to be different” (Dirac 1973, 411). In a discussion at the end of his article on the need for a “drastic revision of our ideas of space and time”, Dirac announced a part of his research agenda that was *independent* of the large number hypothesis:

³¹The qualifications “sign inverted” and “generally accepted” refer to the sign convention which agrees with the coordinate-free definition $Riem(Y, Z)X = \nabla_Y \nabla_Z X - \nabla_Z \nabla_Y X - \nabla_{[Y, Z]} X$. It is preferred in the mathematical literature, including (Weyl 1918b, 5th ed., 131), and also used in the majority of the more recent physics books. The “sign inverted” convention in some of the physics literature goes back to Einstein, e.g. (Einstein 1916, 801), who in turn may have followed Ricci and Levi-Civita. It is found in much of the physics literature from the first half of the 20th century, (Eddington 1923, § 37), (Pauli 1921) up to the influential (Weinberg 1972, eqn. (2.1.3)). Weyl, on the other hand, used the above convention long before the coordinate-free definition of the Riemann tensor was available. It seems to be dominant in the more recent literature on GRT, although Rindler speaks of a 50 % distribution among the two conventions (Rindler 2006, 219).

There is one strong reason in support of the theory. It appears as one of the fundamental principles of Nature that the equations expressing basic laws should be invariant under the widest possible group of transformations ... The passage to Weyl's geometry is a further step in the direction of widening the group of transformations underlying physical laws [in addition to general coordinate transformations, E.S.]. One now has to consider transformations of gauge as well as transformations of curvilinear coordinates and one has to take one's physical laws to be invariant under all these transformations, which imposes stringent conditions on them. (Dirac 1973, 418)

Thus far, Dirac's explanations agreed with the view of C. Brans and R. Dicke. Beyond that similarity, he was following a tendency of the time that sought to exploit possible extensions of the symmetries (automorphisms) of fundamental physics. In contrast to Pauli's insistence on a preferred scale, taken over into the general discourse of JBD theory, Dirac argued that at least three different gauges – the so-called Riemann, Einstein, and “atomic” gauges – had to be considered in different theoretical or observational contexts. He saw no chance of a single preferred gauge. In this respect he clearly differed from Weyl who had speculated that the “atomic” gauge ought to be identical with Weyl gauge.³²

11.3.2.2 Some Remarks on Dirac's Followers

Dirac's proposal for reconsidering Weyl geometry in a modified theory of gravity was taken up by field theorists and a few astronomers. There soon followed an oft-quoted paper by *Vittorio Canuto* and coauthors that gave a broader and more detailed introduction to Dirac's view of Weyl geometry in gravity and field theory (Canuto et al. 1977). It opened with remarks motivating the renewed interest in Weyl geometry based on recent developments in high energy physics: “In recent years, owing to the scaling behavior exhibited in high-energy particle scattering experiments there has been considerable interest in manifestly scale-invariant theories” (Canuto et al. 1977, 1643). This refers to Bjorken scaling and, in particular, the seminal paper (Callan et al. 1970). Still, the authors were careful not to attribute field-theoretic reality to Dirac's scalar function β (Canuto et al. 1977, 1645). Rather they developed modelling consequences for their approach in several different directions: in cosmology, including LNH (large number hypothesis as a gauge condition), for a modification of the Schwarzschild solution in the Dirac framework, and for planetary motion and stellar structure. They concluded by indicating certain heuristic links to gauge fields, which were used in high energy physics during the late 1970s.

³²In his discussion with Einstein Weyl had argued that atomic clocks somehow adapt to the local field constellation via the Weylian scalar curvature (letter to Einstein April 28, 1918 etc (Einstein 1998, vol. 8B, doc. 526, 619)), and similarly in the fourth edition of *Raum - Zeit - Materie* (Weyl 1922, 308f). In the fifth edition Weyl pondered the possibility that the Bohr theory of the atom might give a clue for such an adaptation (Weyl 1918b, 5-th ed., p. 298); compare (Scholz 2017, sec. 5.4).

Canuto was interested in exploring Dirac's idea that the gravitational units of measurement – expressed by a locally dependent parameter of gravity (in place of a constant) and a frequency change for gravitational clocks, like the period of planets orbiting a star – might differ from the atomic units. This assumption would then imply a violation of the strong equivalence principle. In careful evaluations of the astrophysical data available at the beginning of the 1980s, he and his coauthor *Itzhak Goldman* concluded that a tiny difference might well be possible (Canuto/Goldman 1983).

For some years Dirac's approach also attracted the interest of the astronomers *Pierre Bouvier*, *André Maeder* and their coworkers at the Geneva observatory.³³ In November 1977, only a few months after the publication of (Canuto et al. 1977), they submitted a theoretical vindication of “Weyl's geometry as a framework for gravitation” to the journal *Astrophysics and Space Science* (Bouvier/Maeder 1977) as background for a larger research program. Maeder intended to “build some new mechanics” into the framework of Dirac-Weyl-geometrical gravity for purposes of determining the gravitational mass in large systems (clusters, super clusters) on the basis of the virial theorem. His hope was that by means of this “new mechanics” he could account for the *missing mass* identified observationally around the middle of the 1970s by astronomers and astrophysicists (ibid, 341f).³⁴ Initial empirical investigations on the Coma cluster seemed to support Maeder's conjecture (Maeder 1978b,a), but during the years following evidence supporting it dissolved. So the first attempt to bring Dirac's theory to bear in observational cosmology faded away by the early 1980s. We return to this issue in Section 11.6.3.

Unlike empirical astronomers, theoreticians have an open horizon. Dirac's program thus continued to be pursued during the following decades on the theoretical level by, among others, *Nathan Rosen*, working at the Technion in Haifa and the University of Beer Sheva, along with *Mark Israelit*, who immigrated to Israel in 1971 and acquired his Ph.D. at Haifa in 1975. In their continuation of the Dirac program, Rosen and Israelit adopted as far as possible the *e.m.* dogma for a non-conformally coupled scalar field, $k \neq 6$, but with a light massive (Proca-type) photon. Already in his 1982 paper, however, Rosen discussed the possibility of interpreting φ_μ as the potential of a new, hypothetical, heavy massive boson field (see below). During the 1990s, he and Israelit shifted to this latter interpretation, taking it as the preferred physical view of the Weylian scale connection.

Rosen extended Dirac's approach in several respects. He added a scale-invariant mass term \mathcal{L}_m to the Lagrangian, studied the dynamical equations and the corresponding currents and Noether relations, and revisited the question of different gauges (Rosen 1982). Although he recognized the importance of the scale-covariant

³³On the work of Canuto and Maeder, see (Kragh 2006, pp. 126ff)

³⁴For an illuminating historical report on the rise of dark matter, see (Sanders 2010). From a methodological point of view, Maeder's hypothesis was not so far from the later, more pragmatic and more successful approach of modified Newtonian dynamics, MOND, as introduced by *Mordechai Milgrom*.

derivatives corresponding to our (11.3) for giving the Lagrange density a scale-invariant form, he *did not* write the dynamical equations *scale invariantly*. The left hand side of the Einstein equation, e.g., appeared with the Einstein tensor of the Riemannian component of the Weyl metric, ${}_gG = {}_gRiem - \frac{{}_gR}{2}g$ in the notation of our Section 11.2.1, rather than with the respective (scale-invariant) Weyl-geometric tensors $G = Riem - \frac{R}{2}g$. Similarly, the right hand side expressions for the energy-momentum of mass and the scalar field were neither scale covariant nor scale invariant. All terms of the dynamical equations were stripped down to their Riemannian cores. This deprived the Weyl-geometric framework of much of its conceptual strength, even though the equalities were valid in every scale gauge (Rosen 1982, eqn. (121)). This feature remained in all the works of Rosen and Israelit. A scale-co/invariant form for the dynamical equations was only introduced a decade later in the work of Hung Cheng and Drechsler/Tann (Section 11.5.2).

Rosen also hoped to remove the old problem Einstein had raised in 1918 as an objection to Weyl's generalized geometry by introducing what he called a "standard vector" (field) in connection with Dirac's "atomic gauge", in the sense of the Weyl gauge. His aim was to show how this could be made consistent with a non-integrable Weyl-geometric structure. For any timelike vector field u of Weyl weight $w(u) = -1$, the norm $|u| = g(u, u)^{\frac{1}{2}}$ is scale invariant ($w|u| = 2 - 1 - 1 = 0$). If $|u|$ is scale covariantly constant, i.e. $D|u| = 0$ (D the scale-covariant derivative), Rosen called this a *standard vector* field. He then considered the hypothesis that atoms carry a "standard vector" field with them, though he was careful to avoid taking this idea to be a definitive solution of the measurement problem in Weyl-geometric gravity. (Rosen 1982, p. 220f).

Rosen furthermore found that the Noether relations due to the diffeomorphism invariance of the Lagrangian imply the equations of motion for matter, while those induced by scale invariance show that the scalar field equation follows from the Einstein equation and the generalized "electrodynamical" (i.e. scale curvature) equation (Rosen 1982, p. 230). Moreover, by studying Dirac's Lagrangian (11.15) with general coefficient k he realized that for the case of non-conformal coupling the scale curvature equation acquires the form of a generalized Proca equation

$$\nabla_\nu f^{\mu\nu} + m^2\varphi^\mu = 0 \quad (11.17)$$

with $(m [c\hbar])^2 = \frac{1}{2}(6 - k)$ (Rosen 1982, 233). This was consistent with Smolin's observation regarding the Weylian scale connection (cf. Section 11.5.1), which Rosen apparently did not know. He concluded that in the case $k \neq 6$ two physical interpretations for the scale connection were possible: either φ_μ might be understood as representing an electromagnetic field with massive photons of very small mass or else it could be regarded as a "meson" field interacting extremely weakly with ordinary matter.³⁵ Regarding the latter possibility he noted: "These mesons could

³⁵Rosen's "meson" was a hypothetical massive fundamental boson, not a bound state of quarks like the ones in the SM.

conceivably accumulate at the center of galaxies and galaxy clusters and could [provide] the ‘missing mass’ that is needed to give a closed universe.” (Rosen 1982, p. 234) Rosen thus considered an early “dark matter” hypothesis for the Weyl field at a time when the conditions for our present understanding of dark matter in galaxies and structure formation were just forming (Sanders 2010). He mainly related this to the missing mass for cosmological models of positive spatial curvature, while alluding only implicitly to Zwicky’s early observations of a mass problem in galaxy clusters.

For cosmological investigations Rosen also considered a vanishing scale curvature. This led to an integrable Weyl geometry with the logarithm of the Dirac scalar field as the potential of the scale connection, as in (11.9) above (Rosen 1982, eqn. (136)). Although this implies a dynamically trivial extension of Einstein gravity, Rosen found it nevertheless interesting to discuss scaling effects from a geometrical point of view. For a Robertson-Walker type metric $\bar{g}_{\mu\nu}$ of the form

$$d\bar{s}^2 = dt^2 - \frac{a(t)^2}{a_o^2} dl^2 \quad (11.18)$$

he introduced

$$g_{\mu\nu} = \frac{a_o^2}{a(t)^2} \bar{g}_{\mu\nu} .$$

as the *cosmic gauge*. After an appropriate reparametrization of the time coordinate this led to a static Riemannian metric for the model (11.18),

$$g_{\mu\nu} : \quad ds^2 = dT^2 - dl^2 ,$$

with $\beta = \frac{a(t)}{a_o}$ (Rosen 1982, 234ff). He then showed that the cosmological redshift z of a light signal, emitted at time T_o and received at T_1 , remains invariant under rescaling. In the “cosmic gauge” this is no longer due to a spatial expansion of the geometry but rather to the scalar field β , with $z + 1 = \frac{\beta(T_1)}{\beta(T_o)}$. Put equivalently, although Rosen did not mention this, his approach interprets cosmological redshift as due to the Weylian scale connection in the “cosmic gauge” (cf. Section 11.6.2.2).

11.3.2.3 Omote, Utiyama and the Japanese Group

Although unnoticed by Dirac, already in 1971 *M. Omote* in Tokyo had formulated a Lagrangian field theory of gravity with a scale-covariant scalar field coupling

to the Hilbert term. (Omote 1971)³⁶ This was much like the approach in JBD theory, but now explicitly formulated in the framework of Weyl geometry. A little later, more or less at the time of Dirac’s work using Weyl geometry, *Ryoyu Utiyama* (Toyonaka/Osaka) took a similar line, though his main interest was in elementary particle physics.³⁷ In contrast to Dirac, though, he abandoned the *em* dogma and tried to understand the (non-integrable) Weylian scale connection as a new fundamental field. In a series of papers he ventured toward its bosonic interpretation (Utiyama 1973, 1975a,b), and he presented his results at the Seventh International Conference on Gravitation and Relativity (Tel Aviv, June 1974). Utiyama emphasized that a Brans-Dicke field ϕ of weight -1 , imported into Weyl geometry, could serve as a kind of *measure field* (Utiyama’s terminology) with respect to which gauge-invariant measurable quantities could be expressed starting from any gauge (Utiyama 1973, 1975b). This importation made it appear natural that ϕ would be accompanied by a Weylian scale connection φ with non-vanishing curvature (“Weyl’s gauge field”). So Utiyama proposed to explore the ordinary Yang-Mills Lagrangian term for a Weylian scale connection

$$\mathcal{L}_\varphi = -\varepsilon \frac{1}{4} f_{\mu\nu} f^{\mu\nu} \sqrt{|det g|} \quad (\text{here with } \varepsilon = 1) \quad (11.19)$$

(Utiyama 1975b, (2.4)).³⁸ He studied conditions under which “Weyl’s gauge field” admitted plane wave solutions, arriving at the conclusion that such a field would be “tachyonic”, i.e. it would admit perturbations that propagate with superluminal velocities. In Utiyama’s view, this meant that the “boson” had to be confined to the interior of matter particles. Nevertheless he thought that this “unusual field φ_μ might play some role in establishing a model of a stable elementary particle” (Utiyama 1973, 2089).

Utiyama’s results were not generally accepted. *Kenji Hayashi* and *Taichiro Kugo*, two younger colleagues from Tokyo and Kyoto, respectively, reanalyzed his calculations and argued that, with slight adaptations of the other parameters, the sign of ε could be switched, which would lead to a non-tachyonic field. (Hayashi/Kugo 1979, 340f).³⁹ Even so, the scale connection would still have strange physical properties. The two physicists showed, after a careful introduction of Weyl-geometric spinor fields and their Lagrangians (using scale-covariant derivatives), that the scale connection terms canceled. Since they considered only the kinetic term

³⁶This was more than a year before the Trieste symposium at which Dirac talked about his ideas. Apparently the paper remained unknown to Dirac. A second paper by Omote followed after Dirac’s publication and after Utiyama had jumped in (Omote 1974).

³⁷He referred to A. Bregman’s paper, discussed in Section 11.3.3. In Utiyama’s first paper of 1973, Omote went unmentioned; he appears, however, in the references of (Utiyama 1975a).

³⁸Dirac included a similar scale curvature term in his Lagrangian, but did not study its consequences.

³⁹Apparently Hayashi/Kugo used different signature conventions from Utiyama, which resulted in another sign flip in ε .

of fermionic Lagrangians, neglecting the Yukawa term, in this approach neither the scalar ϕ -field nor the scale connection φ coupled effectively to spinor fields.⁴⁰

Thus, at the very moment when a Weylian scale connection φ was interpreted as a “physical” field beyond electromagnetism, this possibility only seemed to puzzle early investigators by posing more riddles than it was able to solve. It failed to couple with matter fields (Hayashi/Kugo) and it looked either “tachyonic” (Utiyama) or, as we shall see below (Smolin, Nieh, Hung Cheng), to be on the order of Planck mass, thus far below anything observable.

11.3.3 Cartan-Weyl-Geometric Approaches

11.3.3.1 The Cartan Geometric Approach to Gravity

Another avenue for Weyl-geometric gravity came from the research tradition initiated by *Dennis Sciama* and *Thomas Kibble*, who developed a theory of gravity by “localizing” the symmetries of Minkowski space, i.e., the Poincaré group (Sciama 1962; Kibble 1961). Their approach treated the “external” symmetries of spacetime in the same manner as the “internal” ones investigated in the gauge theories over Minkowski space in elementary particle physics (isospin, SU_2 , later SU_3 and generalizations), which arose from the works of Yang/Mills and Utiyama. Without explicit recourse to Cartan, they reproduced basic structures of Cartan geometry in field-theoretic terms, but written in classical tensor calculus. The dynamical nature of the infinitesimal translations component of the “localized” Poincaré group found its expression in the asymmetry of the linear connection, viz. torsion. Subsequently, different Lagrangians were investigated and more general groups were studied, in particular the scale-extended Poincaré group or the affine group. In this way a broad field of *gauge theories of gravitation* arose (Blagojević/Hehl 2013).

During the 1970s several authors introduced Cartan geometric methods into this research program; two were particularly prolific: *André Trautman* in Warsaw and *Friedrich Hehl*, from Cologne.⁴¹ They showed that Cartan geometry offered a tailor-made geometric framework for infinitesimalizing (“localizing” in the language of physicists) the symmetries and the currents known from Minkowski space and special relativity. About the same time, the first publications appeared that studied the scale-extended Poincaré group, often called the *Weyl group*,

$$\mathfrak{W} = \mathbf{R}^n \rtimes (SO(1, n - 1) \times \mathbf{R}^+). \quad (11.20)$$

⁴⁰In fact, ϕ and φ were not even coupled to the electromagnetic field, as Hayashi and Shirafuri showed in another paper that same year.

⁴¹(Hehl 1970; Trautman 1972, 1973; Hehl et al. 1976b). For ex-post surveys see (Trautman 2006; Hehl 2017) and the rich reader (Blagojević/Hehl 2013).

In the global view (\mathbf{R}^+, \cdot) operates as the *dilation* group on the translations and, in the case $n = 4$, on the underlying Minkowski space $\mathbf{R}^{(1,3)} \cong \mathfrak{M}/(SO(1, 3) \times \mathbf{R}^+)$. Under localization, or equivalently in the corresponding Cartan space, the infinitesimal groups (Lie algebras) are related in such a way that $so(1, 3) \oplus \mathbf{R}$ operates on the infinitesimal translations, \mathbf{R}^4 . \mathbf{R}^4 is “soldered” in a point-dependent fashion to the tangent spaces of the underlying differentiable space M by specifying a tetrad field, or more generally a *frame field*, i.e., a family of bases of the tangent spaces. In more current mathematical terms, this corresponds to the choice of a *Cartan gauge* in a Cartan space modelled after $\mathfrak{M}/(SO(1, 3) \times \mathbf{R}^+)$, respectively the corresponding Lie algebras (Sharpe 1997). What appears in the global view as an operation on the space itself was thus reshaped, in the infinitesimalized situation, and viewed as a mere change of a Cartan gauge. Weyl’s original ideas from his 1918 geometry as well as those of Dicke and Dirac with regard to unit scaling were well expressed following this approach, while at the same time the setting was extended by introducing translational curvature, *torsion* in Cartan’s terminology.

11.3.3.2 Alexander Bregman

Working in Kyoto, Bregman inferred from (Omote 1971) that localized rescaling could be separated from Weyl’s geometrical interpretation of the infinitesimal length transport. He argued that the point-dependent scale transformations could be treated as “analogous to the introduction of a space-time dependence into the constant parameters of Isospin or Poincaré transformations”. The global scale dimensions d of a physical field X could then be taken over as the “Weyl weight” of X (Bregman’s terminology) in the localized theory (Bregman 1973, p. 668). Bregman first developed a Kibble-like approach to gravity built upon the Poincaré group with tetrad fields h_a^μ , ($a = 0, \dots, 3$ indexing the tetrads, $\mu = 0, \dots, 3$ the coordinates). He introduced a covariant derivative in terms of tetrad coordinates allowing for torsion, and a spin connection expressed by coefficient systems of the form A_{μ}^{mn} with regard to generators S_{mn} of the Lorentz group.⁴² Then he went on “to accommodate” the Weylian scale transformations to the tetrad calculus, in particular by rescaling the tetrads with weight -1 (ibid. pp. 675ff)

$$h_a^\mu \mapsto h'^\mu_a = \Omega^{-1} h_a^\mu. \tag{11.21}$$

This expressed an operation of the scale group on the tetrads, but not on the tangent vectors, which remained unaffected by rescaling.⁴³ It was then necessary to extend the spin connection by a component in the Lie algebra R of the scale group, i.e., a Weylian scale connection $\varphi = \varphi_\mu dx^\mu$,⁴⁴

⁴²Notations have been slightly adapted.

⁴³For $g_{\mu\nu} = h_\mu^a h_{\nu a}$ the convention (11.21) boils down to $g_{\mu\nu} \mapsto g'_{\mu\nu} = \Omega^2 g_{\mu\nu}$.

⁴⁴Bregman, like many other authors, used a sign inverted convention for the scale connection form.

$$\hat{A}_\mu^{mn} = A_\mu^{mn} - (h_\mu^m h^{vn} - h_\mu^n h^{vm})\varphi_v \quad (\text{Bregman 1973, eqn. (3.6)}). \quad (11.22)$$

Bregman remarked that the modified spin connection represented by \hat{A}_μ^{mn} is “Weyl invariant”, and he used it to define an associated scale-covariant derivative $\hat{D}_k = h^\mu_k \hat{D}_\mu$ with the property $\hat{D}_\lambda g_{\mu\nu} = -\varphi_\lambda g_{\mu\nu}$, typical for a Weylian metric like (11.4). The corresponding linear connection $\hat{\Gamma}_{\mu\nu}^\lambda = \Gamma_{\mu\nu}^\lambda + \delta_\mu^\lambda \varphi_\nu + \delta_\nu^\lambda \varphi_\mu - g_{\mu\nu} \varphi^\lambda$ generalized the Weylian affine connection, but was no longer symmetric; rather it included the scale-invariant *torsion* tensor

$$T_{\mu\nu}^\lambda = \hat{\Gamma}_{\nu\mu}^\lambda - \hat{\Gamma}_{\mu\nu}^\lambda = \Gamma_{\nu\mu}^\lambda - \Gamma_{\mu\nu}^\lambda. \quad (11.23)$$

In retrospect, one can see how Bregman’s paper could be reworked using the symbolism for a Cartan space modelled after the homogeneous space $\mathfrak{W}/(SO(1, 3) \times \mathbf{R}^+)$, later called a *Cartan-Weyl space* (or the other way round).⁴⁵ This terminology was not Bregman’s; he used Cartan’s geometric language rather parsimoniously, thus only with regard to the underlying Riemann-Cartan structure, modelled after $\mathfrak{P}/SO(1, 3)$ with $\mathfrak{P} = \mathbf{R}^4 \rtimes SO(1, 3)$ denoting the Poincaré group. He did not think in geometric terms about the extension of this structure by rescaling the tetrads, the physical (spinor, vector etc.) fields, or the associated spin connection, etc.

Bregman was more interested in showing how to form Weyl-invariant Lagrangians L from matter Lagrangians L^M (his notation) of scale dimension -4 with regard to “constant parameter scale transformations”. He noted that many Lagrange densities are also invariant under all conformal transformations of the Minkowski space, including the special conformal ones (“in particular this is generally true of theories whose quantized versions are renormalizable”), to which he added: “In our case such a wider invariance of L^M implies in turn that the Poincaré gauge invariant lagrangian L^P is already Weyl invariant with $L = L^P$ ” (Bregman 1973, p. 678). This cogent remark in effect generalized Pauli’s observation that a massless Dirac-spinor field is invariant under Weyl transformations without assuming a coupling to a Weylian scale connection (Pauli 1940).

Finally, Bregman gave a short discussion of an integrable scale connection with potential σ , $\varphi_\mu = -\partial_\mu \sigma$. He considered σ as an “independent dynamical variable” which is “connected to the translation or spin gauge fields” only through the field equations (ibid. p. 687). This approach facilitated the building of Weyl-invariant Lagrange densities. As an example he presented a Lagrangian, which was similar to Omote’s (and Dirac’s not yet published one), by including an additional torsion

⁴⁵Cf. (Sharpe 1997).

term (Bregman 1973, eqn. (5.2)).⁴⁶ All in all, this was a remarkable paper which seems to have been largely neglected in the following development.

11.3.3.3 Charap/Tait

About a year later, *John Charap* and *W. Tait*, both in London, presented a “gauge theory of the Weyl group” by building on earlier work of (Yang/Mills 1954; Utiyama 1956; Kibble 1961; Dirac 1973), whereas Bregman’s paper went unnoticed. They introduced the Weyl group as the “simplest possible non-trivial enlargement of the Poincaré group” (Charap/Tait 1974, p. 250), where by “non-trivial” they were apparently hinting at the semidirect product operation of \mathbf{R}^+ on the translations. Like Bregman before them, they explored “the consequences of demanding for a theory of matter fields that it be invariant under the transformations of the Weyl group” (ibid.).

They began by studying the infinitesimal Weyl transformations on Minkowski space endowed with a globally Weyl invariant Lagrangian $L(\chi, \chi')$, where L depends on coupled fields χ and their first derivatives (indicated by χ'). They derived the Noether relations with regard to translations, rotations, and dilations, though without mentioning Noether.⁴⁷ If the Euler-Lagrange equations for all fields are satisfied (“on shell”), then conservation laws for expressions corresponding to the symmetries follow (Noether’s first theorem). Most of them could easily be identified with well-known physical quantities and were called *canonical currents*: the canonical energy momentum current ${}_cT_v^\mu$, the canonical angular momentum current $M_{v\lambda}^\mu$, and additionally a *canonical dilation current* ${}_c\Delta^\mu$. The last of these evaded any immediate physical interpretation. But in analogy with the angular momentum, which can be decomposed into an internal (spin) contribution of the fields and an external, orbital component, the dilation current could be decomposed into

$${}_c\Delta^\mu = J^\mu + {}_cT_v^\mu x^v, \quad J^\mu = \frac{\partial L}{\partial(\partial_\mu \chi)} w(\chi) \chi, \tag{11.24}$$

with an internal component J^μ (summation of the field components understood to be included) and an external one depending on the origin of the coordinate system in Minkowski space. The external component was, of course, due to the dilational operation of \mathbf{R}^+ on the underlying space. Of interest here is what happens under “localization” of the symmetries.

⁴⁶The torsion term $\frac{2}{3}T_{\mu\alpha}^\alpha \partial^\mu \sigma^2$ in Bregman’s equation is not scale invariant by itself, but his entire Lagrangian density is scale invariant.

⁴⁷This was characteristic for the time. Over several decades the knowledge about the invariance properties of Lagrangian field theories and the know-how of dealing with it spread with marginal or no reference at all to Noether’s seminal paper (Noether 1918), cf. (Kosmann-Schwarzbach 2011).

Charap and Tait localized following the approach of Kibble, i.e., they made the group operations point dependent by introducing a tetrad field h_k^μ and *gauge fields* (geometrically connections), the first one with values in the Lorentz algebra, given by coefficients A_μ^{ij} with respect to the algebra generators S_{ij} , the other taking values in the real numbers, given by a system of A_μ . This allowed them to define the scale-covariant derivative as

$$D_\mu \chi = \partial_\mu \chi + \frac{1}{2} A_\mu^{ij} S_{ij} \chi - A_\mu w(\chi) \chi \quad (\text{Charap/Tait 1974, eqn. (3.7)}). \tag{11.25}$$

Like Bregman, they considered Lagrange densities \mathfrak{L} constructed from globally invariant Weylian ones, L , obtained by substituting scale-covariant derivatives for partial derivatives (“minimal coupling” in the later physics idiom) and by using the volume element $|h|^{-1} dx$ with $|h| = \det(h_k^\mu)$. They analyzed the resulting gauge field dynamics and considered their sources, usually the “right hand side” of the equations, as “modified currents”.⁴⁸

Among the localized *canonical* (Noether) currents they explicitly mentioned only the one for *energy-momentum*. In analogy with the global case, they defined

$${}_c \mathfrak{T}_\mu^\lambda = \frac{\partial \mathfrak{L}}{\partial (\partial_\lambda \chi)} \partial_\mu \chi - \delta_\mu^\lambda \mathfrak{L}, \tag{11.26}$$

while commenting on its difference from the *dynamical energy momentum* arising from variational derivation of the matter Lagrangian with respect to the gravitational field. The latter, written with respect to a coordinate basis as a “world tensor” \mathfrak{T}_μ^λ , can be derived from the canonical energy momentum ${}_c \mathfrak{T}_\mu^\lambda$ by adding terms expressed using dynamical spin and dilation quantities.⁴⁹ The authors made no mention of the corresponding conservation theorem or the other Noether currents, not even the canonical dilation current which was around the corner,

$$c \Delta^\mu = - \frac{\partial \mathfrak{L}}{\partial (\partial_\mu \chi)} w(\chi) \chi. \tag{11.27}$$

In a separate passage on the “geometrical interpretation” of their theory, they explained how their approach took up Weyl’s scale geometry of 1918 and how they had generalized it by including torsion; but Cartan geometry was neither mentioned nor used.

⁴⁸ ${}_c \mathfrak{T}_\mu^k = \frac{\partial \mathfrak{L}}{\partial h_k^\mu}$, $\mathfrak{S}_{ij}^\mu = -2 \frac{\partial \mathfrak{L}}{\partial A_\mu^{ij}}$, $\mathfrak{D}^\mu = \frac{\partial \mathfrak{L}}{\partial A_\mu}$ (Charap/Tait 1974, p. 256, notation slightly changed). Compare the *dynamical currents* by Hehl et al. below (11.29). For the dynamical matter energy current variational and partial derivatives usually coincide. Charap and Tait may have (wrongly) generalized this property to the other currents.

⁴⁹ $\mathfrak{T}_\mu^\lambda = {}_c \mathfrak{T}_\mu^\lambda - \frac{1}{2} \mathfrak{S}_{ij}^\lambda A_\mu^{ij} - \mathfrak{D}^\lambda A_\mu$ (Charap/Tait 1974, eqn. (3.22)).

11.3.3.4 Hehl et al. at the Kiel Conference

During the next decade several papers appeared that dealt with gravity in a Cartan-Weyl setting, among them (Kasuya 1975; Obukhov 1982; Nieh 1982).⁵⁰ Not all of these contained new insights, nor were their statements always reliable; but they indicate a slow broadening of interest in Weyl-geometric gravity with torsion. This interest acquired a wider context with the onset of Cartan geometric *metric-affine* studies of gravity. A general discussion of this research would go far beyond the limits of this survey,⁵¹ but at least one of these papers has to be commented on here: (Hehl et al. 1988c). This paper was originally presented by *Friedrich Hehl* and coauthors at the Weyl centenary conference held in Kiel in 1985. Its discussion of gravity based on Cartan's metric-affine geometry, with particular emphasis on the Weyl group $\mathfrak{W} \subset \mathbf{R}^n \rtimes GL(n, \mathbf{R})$, may be taken as paradigmatic for how Weyl-geometric aspects were dealt with in this research program.

In modernized language, this theory uses a Cartan geometry modelled on the Klein space $\mathfrak{A}/GL(n, \mathbf{R})$ with the affine group $\mathfrak{A} = \mathbf{R}^n \rtimes GL(n, \mathbf{R})$ and corresponding Lie algebras. Additionally, it assumes an independently given (Lorentzian) metric g . The local description again involved an n -frame field h_a^μ ($a, \mu = 0, \dots, n-1$) and its dual system of forms h_μ^a characterizing a Cartan gauge or the translational connection. It also utilizes a connection with values in the Lie algebra of the isotropy group $gl(n, \mathbf{R})$ (generalizing the rotations of Cartan-Riemann geometry) given by the coefficient system $A^b_{a\mu}$, along with a metric of Lorentzian signature given by $g_{\mu\nu} = h_\mu^a h_\nu^b g_{ab}$. Let the corresponding covariant derivative be denoted by D_a , respectively D_μ if transcribed to coordinate indices. Of course, in general the derivative of the metric does not vanish, $D_\lambda g_{\mu\nu} = -Q_{\lambda\mu\nu}$, where $Q_{\lambda\mu\nu}$ is usually called the *non-metricity* of the derivative, which is symmetric in its last two indices. With respect to coordinate bases the linear connection is $\Gamma_{\mu\nu}^\lambda = h_\mu^a h_\nu^b A^{\lambda}_{ab}$.

In case Γ is symmetric, it can be decomposed as $\Gamma = {}_g\Gamma + {}_Q\Gamma$, with a Riemannian (Levi-Civita) component ${}_g\Gamma$ and a component due to the non-metricity ${}_Q\Gamma$.⁵² $Q_{\lambda\mu\nu}$ can also be decomposed into a traceless part $\mathcal{Q}_{\lambda\mu\nu}$, with respect to the last two indices, and its trace q_λ . Hehl and his coauthors called the trace part q_λ of the non-metricity the “celebrated Weyl vector”

$$q_\lambda = \frac{1}{n} Q_{\lambda\nu}{}^\nu \quad (\text{Hehl et al. 1988c, p. 252}). \quad (11.28)$$

⁵⁰All of them are mentioned in (Blagojević/Hehl 2013, chap. 8).

⁵¹In particular F. Hehl and his various coauthors were active in this field, (Hehl et al. 1976a, 1988a,b, 1989, 1995). More papers by other authors, though only a few of those just mentioned, appear in (Blagojević/Hehl 2013, chap. 9).

⁵² $Q_{\mu\nu}^\lambda = \frac{1}{2}(Q_{\mu\nu}{}^\lambda - Q_{\nu\mu}^\lambda + Q_\nu{}^\lambda{}_\mu)$

In our notation $\varphi_\lambda = \frac{1}{2}q_\lambda$ is a *part of the connection* and can also be expressed by $\varphi_\lambda = \frac{1}{n}A^a_{a\lambda}$. In the case of vanishing \mathcal{Q} , the “Weyl vector” satisfies the metric compatibility condition of Weyl geometry, our eqn. (11.4). In this situation, the linear connection Γ will coincide with the Weylian invariant affine connection, allowing for an embedding of Weyl geometry in the wider framework of metric-affine geometry.

In this framework, the data $(h^\mu_a, A^b_{a\mu}, g_{ab})$ were considered as dynamically independent field components of an extended theory of gravity (Hehl et al. 1988c, sec. 6, notation changed). For a set of matter fields Ψ and a matter Lagrangian \mathcal{L}_m , minimally coupled to the affine-metric structure $\mathcal{L}_m(g_{ab}, h^\mu_a, \Psi, D_a\Psi)$, the authors defined the *dynamical matter hypermomentum current* of their generalized theory by the variational derivative with regard to the full isotropy connection,

$$\Delta_b^{a\mu} = |h| \frac{\delta \mathcal{L}_m}{\delta A^b_{a\mu}}, \quad \text{with } |h| = \det(h^\mu_a), \quad (11.29)$$

and decomposed it into its rotational, dilational, and shear components (Hehl et al. 1988c, p. 274f).

In the context of their lecture in Kiel, the dynamical *dilational current*

$$\Delta^\mu := \Delta_a^{a\mu} = |h| \frac{\delta \mathcal{L}_m}{\delta A^a_{a\mu}} = \frac{|h|}{n} \frac{\delta \mathcal{L}_m}{\delta \varphi_\mu} \quad (11.30)$$

was of particular importance. The authors conceded that the shear current was “remote from physical experience”; but for the dilational current they saw “supporting evidence” in Bjorken scaling of deep inelastic scattering experiments and, on the theoretical level, in certain models of supergravity (Hehl et al. 1988c, pp. 244, 275). This was a central point for their talk. Already in their abstract they announced: “In the light of modern developments in particle physics, this coupling of the Weyl vector to the material dilation current is an unalterable part in any viable theory of a general-relativistic type, which comprises a Weylian piece” (Hehl et al. 1988c, p. 241). This thesis stood in contrast to the observation of Bregman (who was cited by these authors) that for matter fields with conformally invariant Lagrangians there was no need to assume any coupling to the scale connection (the “Weyl vector”). Still, Hehl and his coauthors were able to give an example of a scalar field with non-vanishing dilational current (see below).

They also analyzed the dynamics of the metric affine theory on a quite general level, though, at first, with no particular Lagrangian specified. They described the form of the three dynamical equations corresponding to the decomposition of the general linear group into its shear, rotational, and dilational components, showing that only two of these were dynamically independent due to the results of the second Noether theorem (Hehl et al. 1988c, sec. 8). A short discussion of specific Lagrangians followed, drawing consequences from a scale-covariant “primordial scalar field, the so-called *dilaton field* $\sigma(x)$ ” (Hehl et al. 1988c, p. 282,

emph. in original) of Weyl weight $w(\sigma) = -\frac{n-2}{2}$ and the usual scale-invariant quadratic kinetic term

$$\mathcal{L}_\sigma = \frac{|h|}{2} D_\mu \sigma D^\mu \sigma. \quad (11.31)$$

As a “primordial” field, σ was considered to be part of the matter sector, but because of the scale-covariant derivative in \mathcal{L}_σ it contributed to the matter dilational current with

$$\Delta^\mu(\sigma) = \frac{n-2}{2n} \sigma D^\mu \sigma \quad (\text{Hehl et al. 1988c, eqn. (10.11)}). \quad (11.32)$$

The authors warned not expect easy empirical support from laboratory experiments since: “Local scale invariance of fundamental interactions is expected to be valid only approximately in the high energy limit of Bjorken scaling or exactly at the onset of the big bang” (ibid, p. 285). They assumed a breaking of scale symmetry down to the Poincaré group “after a very short time lag” (to the big bang) and proposed a quartic potential for σ with a symmetry breaking quadratic term similar to the Higgs potential. In the end, their gravitational Lagrangian *boiled down to Einstein gravity* with cosmological constant “plus some supplementary terms known from Poincaré gauge theory” (ibid, p. 242). By the “supplementary terms” they apparently referred to the torsion-spin coupling which arises in Einstein-Cartan gravity. This leads to a serious modification of Einstein gravity only at extremely high mass densities.⁵³

In the framework of Hehl et al., Weyl-geometric modifications of gravity were to be expected only under conditions that were even more extreme than those for torsion. In their view, Weyl-geometric effects seemed to be banned to a speculative realm close to the “big bang”, one of the great adventure playgrounds of late 20th century theoretical physics. In Sections 11.5 and 11.6 we shall see that this need not necessarily be so, if other perspectives are taken into account. But before we turn to these investigations, we must first consider another road that led to a revival of Weyl’s scale geometry. This path arose from attempts to geometrize the dynamics of non-relativistic quantum mechanics, a development with very different roots from those that led to the Omote-Dirac-Utiyama and the Cartan geometric approaches discussed in this section.

⁵³ According to later estimates the torsion-spin coupling of Einstein-Cartan gravity becomes important only close to 10^{38} times the density of a neutron star, which signifies energy densities at the hypothetical grand unification scale of elementary particle interactions (Trautman 2006, p. 194), (Blagojević/Hehl 2013, p. 108). Moreover, there seems to be a problem in defining the spin density of gauge particles, in particular of gluons. This would lead to leaving a large part of the nucleon spin underdetermined (Kleinert 2008, sec. 20.2) (I thank H. Ohanian for this information).

11.4 Weyl's Scale Connection a Geometrical Clue to Quantum Mechanics?

11.4.1 Bohmian Mechanics as a Background

11.4.1.1 Bohm's "Causal" Approach to QM

In the early 1950s *David Bohm* proposed an alternative approach to non-relativistic quantum mechanics (QM) based on an often discussed heterodox "causal" interpretation of quantum phenomena⁵⁴ His core idea was to reintroduce exact particle trajectories into the description of quantum systems, which were taken to be guided by a pilot wave evolving according to the Schrödinger equation of ordinary quantum mechanics. In effect, he took up earlier ideas of *Louis de Broglie* on the dualism of wave and particle aspects in QM, which had been critically debated in the late 1920s. This approach was mathematically close to a hydrodynamic picture of the Schrödinger equation, first considered by *Ernst Madelung* in 1926. Madelung noticed that his "hydrodynamical" current was subject to a non-classical term which could be interpreted as a kind of force function due to the " 'inner forces' of the continuum" (Madelung 1926, p. 323). Bohm extended these older ideas, among others, by an analysis of the measuring process.⁵⁵ He could thus avoid stipulating a "collapse" of the wave function, which was usually assumed for extracting real valued measuring values from the observables given by the Hermitian operators of QM (Bacciagaluppi 2009, chap. 11). Bohm wanted to challenge the mainstream ("Copenhagen") interpretation of QM which he accepted as consistent but still regarded as unsatisfactory from a foundational and natural philosophic point of view. His goal was to find an *alternative interpretation* of QM which did not affect the dynamics, at least not "in the domain of dimensions of the order of 10^{-13} cm". This ought to make it possible "to conceive of each individual system as being in a precisely definable state, whose changes with time are determined by definite laws, analogous to (but not identical with) the classical equations of motion" (Bohm 1952a, p. 167).

Bohm started by considering any Schrödinger equation for a wave function $\psi(x) = a(x)e^{iS(x)}$ ($x \in \mathbf{R}^3$) governed, e.g. in the case of a single particle of mass m with an external potential $V(x)$, by the equation

$$i\hbar \frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2m} \nabla^2 \psi + V(x)\psi. \quad (11.33)$$

⁵⁴(Bohm 1952a,b)

⁵⁵Bohm realized the kinship of his approach to the earlier proposals of de Broglie only after he had finished his manuscript (Bohm 1952a, p. 167). In a footnote added in proof he also referred to Madelung's "similar" approach of 1926, adding the remark "... but like de Broglie he did not carry this interpretation to a logical conclusion" (ibid.).

From this one can associate a pair of coupled differential equations for the phase $S(x)$ of $\psi(x)$ and the probability density $\rho(x) = a(x)^2$:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \left(\rho \frac{\nabla S}{m} \right) = 0, \quad (11.34)$$

$$\frac{\partial S}{\partial t} + H(x, t) = 0, \quad (11.35)$$

$$\text{where } H(x, t) = \frac{(\nabla S)^2}{2m} + V(x) - \frac{\hbar^2}{4m} \left(\frac{\nabla^2 \rho}{\rho} - \frac{1}{2} \frac{(\nabla \rho)^2}{\rho^2} \right)$$

(Bohm 1952a, eqs. (6), (7)). Equation (11.35) has the form of a *Hamilton-Jacobi* equation for a point particle with the principal function S and conjugate momenta $p_k = \partial_k S$, equivalently the velocity $v = \frac{\nabla S}{m}$. The total potential $\tilde{V}(x) = V(x) + U(x)$ deviates from the classical $V(x)$ by

$$U(x) = -\frac{\hbar^2}{4m} \left(\frac{\nabla^2 \rho}{\rho} - \frac{1}{2} \frac{(\nabla \rho)^2}{\rho^2} \right) = -\frac{\hbar^2}{2m} \frac{\nabla^2 a}{a}. \quad (11.36)$$

Bohm considered $U(x)$ as a kind of *quantum potential* added to the classical one. The trajectories of the Hamilton-Jacobi system have velocities $v = \frac{\nabla S}{m}$ normal to the level surfaces given by constant values of S . Thus (11.34) acquires the form of a continuity equation for an ensemble of point particles following the family of trajectories with the density ρ :

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho v) = 0. \quad (11.37)$$

Bohm argued that this might be “the nucleus of an alternative interpretation for Schroedinger’s equation” (Bohm 1952a, p. 170). Madelung had already obtained a pair of equations similar to (11.35, 11.37) in coming up with his hydrodynamical interpretation, about which he was more cautious. He did not consider his equations as more fundamental than Schrödinger’s wave mechanics, but rather saw them as a “model representation” from which one could derive the essential features of the latter (Madelung 1926).⁵⁶

In Bohm’s alternative interpretation of the equations a quantum particle might seem to no longer be subject to Heisenberg indeterminacy, since it appears to follow a specified trajectory of the system (11.35). However, the Heisenberg principle was implicitly upheld because only a probability satisfying (11.37) could be assigned to a trajectory passing through specified regions in the level surfaces. Via the quantum potential (11.36) a wave function satisfying (11.33) would operate as a non-local

⁵⁶“Die hydrodynamischen Gleichungen sind also gleichwertig mit denen von Schrödinger und liefern alles, was jene geben, d. h. sie sind hinreichend, um die wesentlichen Momente der Quantentheorie der Atome modellmäßig darzustellen” (Madelung 1926, p. 325).

guiding structure, a “pilot wave” (though this terminology was not used by Bohm) for the motion of a quantum particle. This was quite close to the spirit of de Broglie’s theory from the 1920s (Bacciagaluppi 2009).

Although Bohm extended de Broglie’s and Madelung’s view by an analysis of the measuring process, his proposal did not meet with an immediate positive response in the quantum physics community (Myvold 2003). It was only in the longer run that various authors took it up and pursued programs along these lines, sometimes with a quite different outlook on the underlying ontology or by employing new mathematical ideas. Despite these differences, such investigations can be seen as belonging to a common family of *de Broglie – Madelung – Bohm (dBMB)* approaches.⁵⁷

Important for our context was the stepwise extension of the Bohmian approach to relativistic quantum mechanics, in particular for *Klein–Gordon* particles given by a complex field of spin zero, $\psi(x) = a(x)e^{i\frac{S}{\hbar}}$. This pseudo-classical field takes values in the complex numbers, like a Schrödinger wave function, but lives on the Minkowski space with $x = (x^0, \dots, x^3)$. Moreover, $\psi(x)$ involves a more intricate interpretation than does Born’s probability rule. In case of an electromagnetic interaction with potential A_μ , the field of a Klein-Gordon particle of mass m and charge e satisfies the dynamical equation

$$\left(\frac{\hbar}{i}\partial_\mu - \frac{e}{c}A_\mu\right)\left(\frac{\hbar}{i}\partial^\mu - \frac{e}{c}A^\mu\right)\psi = (mc)^2\psi \quad (11.38)$$

in the signature $(+---)$ of the Minkowski space ($\partial_o = c^{-1}\partial_t$). Here the Bohmian method leads to the Hamilton-Jacobi and continuity equations

$$\left(\partial_\mu S - \frac{e}{c}A_\mu\right)\left(\partial^\mu S - \frac{e}{c}A^\mu\right) = m^2c^2 + \hbar^2 Q, \quad (11.39)$$

$$\partial_\mu(a^2\partial^\mu(S - \frac{e}{c}A_\mu)) = 0, \quad (11.40)$$

with a “quantum term” similar to (11.36),⁵⁸

$$Q = \frac{\partial_\mu\partial^\mu a}{a}. \quad (11.41)$$

⁵⁷(Passon 2015, 2004; Dürr et al. 2009), I thank O. Passon for his helpful explanations of Bohmian mechanics. For relativistic generalizations, see, e.g., (Nicolic 2005). For recent critical remarks see Chen (2016).

⁵⁸Cf. (Nicolic 2005, p. 554) for vanishing em potential.

11.4.1.2 A Geometrization Idea Due to de Broglie

In his later years, de Broglie joined in pursuing this research program by pondering the connection between the Jacobi flow of the Klein-Gordon field and general relativity (de Broglie 1960, pp. 118ff). He forwarded a hypothesis according to which quantum particles are constituted by extremely dense tiny regions of the field, governed by unknown non-linear equations, while in the exterior of these regions the known linear equations of quantum mechanics hold. He called these regions “singular” and investigated whether the motion of such “singular” regions might follow geodesics in a manner similar to the motion of singular regions in general relativity, a problem studied by Einstein and Grommer in the 1920s.⁵⁹ In this context he considered the right hand side of (11.39) as a kind of “variable rest mass” which had to be calculated in the “immediate vicinity of the particle” (de Broglie 1960, p. 116):

$$\mathfrak{M}_o = \sqrt{m_o^2 + \frac{\hbar^2}{c^2} \frac{\partial \nu \partial^\nu a}{a}} \quad (11.42)$$

Some authors would later call \mathfrak{M}_o the “quantum mass” of a Klein-Gordon field (see Section 11.4.3).

De Broglie considered the trajectories $x^\mu(s)$ of the Hamilton-Jacobi flow of a particle “in the absence of electromagnetic and gravitational fields” (ibid. p. 119, emph. in original) with 4-velocity, $u^\mu = \frac{d}{ds}x^\mu$, normalized to $u_\mu u^\mu = 1$,

$$u^\mu = (\mathfrak{M}_o)^{-1} \partial^\mu S. \quad (11.43)$$

He then found that these trajectories satisfy the geodesic equations of a metric $g_{\mu\nu}$ arising from the Minkowski metric $\eta_{\mu\nu}$ by conformal rescaling

$$g_{\mu\nu} = \frac{\mathfrak{M}_o^2}{m_o^2} \eta_{\mu\nu}. \quad (11.44)$$

From this he concluded that “even if the particle is not subjected to any gravitational or electromagnetic field, its possible trajectories (...) are the same as if space-time possessed non-Euclidean metrics defined by $[g_{\mu\nu}]$ ” (de Broglie 1960, p. 120).

This was an interesting geometrization argument, and de Broglie would not remain the only one to ponder a connection between the framework of dBMB quantum mechanics and general relativity. Here, though, we are mainly interested in later authors who tried to make progress by attempting a geometrization of QM using the framework of Weyl geometry.

⁵⁹According to de Broglie it was J.-P. Vigi er who made him aware of a parallel between his hypothesis and the work of Einstein and Grommer (de Broglie 1960, p. 92).

11.4.2 *Santamato's Proposal for Geometrizing Quantum Mechanics*

11.4.2.1 Two Phases of Work on the Program

In the 1980s *Enrico Santamato* in Naples proposed a new approach to quantum mechanics (Santamato 1984a,b, 1985). It was based on studying weak random processes of ensembles of point particles moving in a Weylian modified configuration space. He compared his approach to that of *Madelung-Bohm* and to the stochastic program of *Feynès-Nelson*.⁶⁰ While the latter dealt with stochastic (Brownian) processes, Santamato's approach was closer to the view of Madelung and Bohm because it assumed only random initial conditions, with classical trajectories given in Hamilton-Jacobi form (this explains the attribute "weak" above). One can interpret Bohm's particle trajectories as deviating from those expected in Newtonian mechanics by some "quantum force". Santamato found this an intriguing idea, but he deplored its "mysterious nature" which "prevents carrying out a natural and acceptable theory along this line". He hoped to find a *rational explanation* for the effects of the "quantum force" by means of a *geometry* with a modified affine connection of the system's configuration space. Then the deviation from classical mechanics would appear as the outcome of "fundamental properties of space" (Santamato 1984a, p. 216), understood in the sense of *configuration space*.

In his first paper Santamato started from a configuration space with coordinates (q^1, \dots, q^n) endowed with a Euclidean metric. More generally, his approach allowed for a general positive definite metric g_{ij} , and later even a metric of indefinite signature for dealing with general coordinates of n -particle systems, and perhaps, in a further extension, with spin. The Lagrangian of the system, and the corresponding Hamilton-Jacobi equation, contained the metric, either explicitly or implicitly. This Euclidean, or more generally Riemannian, basic structure was complemented by a Weylian scale connection. Santamato's central idea was that the modification of the Hamilton-Jacobi equation induced by a properly determined *scale connection* could be used to express the *quantum modification* of the classical Hamiltonian, much like was done in the Madelung-Bohm approach. Then the quantum aspects of the systems would be geometrized in terms of Weyl geometry, surely a striking and even beautiful idea, if it should work.

Santamato thus headed towards a new program of *geometrical quantization sui generis*. This had nothing to do with the better known geometric quantization program initiated more than a decade earlier by J.-N. Souriau, B. Kostant and others, which was already well under way in the 1980s (Souriau 1966; Kostant 1970; Simms 1978). In the latter, geometrical methods underlying the canonical quantization were studied. Starting from a *symplectic phase space* manifold of a classical system, the observables were "pre-quantized" in a Hermitian line bundle,

⁶⁰For E. Nelson's program to re-derive the quantum dynamics from classical stochastic processes and classical probability see (Bacciagaluppi 2005).

and finally the Hilbert space representation of quantum mechanics was constructed on this basis.⁶¹ Santamato's geometrization was built upon a different structure, Weyl geometry rather than symplectic geometry, and had rather different goals.

Like other proposals in the dBMB (de Broglie-Madelung-Bohm) family, Santamato's program encountered little positive response. In the following decades he shifted his research to more empirically based studies in nonlinear optics of liquid crystals and quantum optics. Perhaps a critical paper by *Carlos Castro Perelman*, a younger colleague who knew the program nearly from its beginnings, contributed to what would be an extended period of interruption. Castro discussed "a series of technical points" which seemed important for Santamato's program from the physical point of view (Castro 1992, p. 872).⁶² Among the problems he mentioned were several of a more foundational nature beyond those of more technical import: (i) how to specify the random initial data for an ensemble of particle paths, (ii) the basis for the Hilbert space interpretation of the theory, and (iii) the relationship of Santamato's approach to the Feynman path integral quantization. He also criticized the lack of a rigorous hypothesis in the choice of the particle's Lagrangian, the non-definite character of the probability density in the case of a Klein-Gordon particle (which could appear if the foliation with respect to the principal function S is not timelike), the dependence of the particle's effective mass on the Weylian scalar curvature (in the configuration space), and some other more technical points.

After the turn of the century, Santamato came back to foundational questions, working closely together with his colleague *Francesco De Martini* from the University of Rome. Both had cooperated in their work on quantum optics already for many years. In the 2010s they turned to geometrical quantization in a series of joint publications that continued the program Santamato had started three decades earlier. They showed how to deal with spinor fields in this framework, in particular with the Dirac equation (Santamato/DeMartini 2013), and they discussed the famous *Einstein-Podolsky-Rosen* (EPR) non-locality question (De Martini/Santamato 2014a,b). Moreover, they analyzed the helicity of elementary particles and showed that the *spin-statistics relationship* of relativistic quantum mechanics could be derived in their framework without invoking arguments from quantum field theory (De Martini/Santamato 2015, 2014c, 2016). In this new series of papers, they took Minkowski space as the starting point for their construction of the configuration spaces, which could be extended by internal degrees of freedom. Moreover, they enlarged the perspective by making a transition from point

⁶¹ See, e.g., (Woodhouse 1991) or (Hall 2001, chaps. 22/23). A classical monograph on the symplectic approach to *classical* mechanics is (Abraham/Marsden 1978); but this does not discuss spinning particles. In the 1980s the symplectic approach was already used as a starting platform for (pre-)quantization to which proper quantization procedures could then hook up, see e.g. (Śniatycki 1980). Souriau was an early advocate of this program. In his book he discussed relativistic particles with spin (Souriau 1970, §14).

⁶² Carlos Castro later added his mother's name Perelman to his second name. Under this name he is mentioned in the acknowledgements of (Santamato 1984b). At that time he was a research assistant at the University of Texas, Austin, where he acquired his Ph.D. in 1991.

dynamical Lagrangians to a dynamically equivalent description in terms of scale-invariant field-theoretic Lagrangians in two scalar fields (see below).

It is unnecessary to go further into details about these often quite technical articles; here I only want to concentrate on the basic question of the geometrization program.⁶³ For this, it suffices to see how Santamato’s intriguing idea of introducing a Weyl-geometric structure on the configuration space can be used to model the Bohmian effects of quantum systems.

11.4.2.2 The Geometrization of the Configuration Space

Summing up, Santamato’s idea was to consider dynamical systems with finite degrees of freedom, parametrized by a configuration space V with parameters q^1, \dots, q^n , which is endowed with a pseudo-Riemannian metric g_{ij} of arbitrary signature (Santamato 1984a). In the case of a non-relativistic k -particle system without inner degrees of freedom, this could be the product of Euclidean 3-metrics, for relativistic particles in Minkowski space with metric $\eta = \text{diag}(-1, 1, 1, 1)$ the signature was $(3k, 3)$ (Santamato 1984b).

In the case of a relativistic 1-particle system with spin, the product of the Minkowski space \mathbf{M} and the Lorentz group served as configuration space $V = \mathbf{M} \times SO(3, 1)$, where the second factor parametrizes “hidden” rotational degrees of freedom of the particle (Santamato/DeMartini 2013, p. 634). By an astute choice of coordinates for V , $(q^1, \dots, q^{10}) = (x^\mu, \theta^\alpha)$ (with $\alpha = 1, \dots, 6$) with generalized “Euler angles” θ^α for parametrizing $SO(3, 1)$, the authors introduced a metric (g_{ij}) by a block matrix composed of the Minkowski metric $\eta_{\mu\nu}$ and a “metric of the parameter space of the Lorentz group” $g_{\alpha\beta}$ with signature $(+++--)$. A frame given by e_a^μ can be characterized by the Lorentz transformation θ that takes the standard basis into the given one, which may now be written as $e_a^\mu(\theta)$. The metric on the Lorentz group component was derived from the group operation on the frames by measuring the Minkowski squared norm induced by infinitesimal rotations of the Euler angles (summation over all frame vectors):

$$g_{\alpha\beta}(\theta) = -a^2 \eta_{\mu\rho} \eta_{\nu\sigma} \omega_\alpha^{\mu\nu}(\theta) \omega_\beta^{\rho\sigma}(\theta) \quad (11.45)$$

with $\omega_\alpha^{\mu\nu}(\theta) = g^{\rho\nu} e_\rho^a(\theta) \frac{\partial}{\partial \theta^\alpha} e_a^\mu(\theta)$.

The factor a^2 was only made explicit elsewhere by the authors and in passing (De Martini/Santamato 2014a, p. 3313). Here a , which expresses the *gyromagnetic*

⁶³Santamato and De Martini promoted the geometrical side of their research under different headings: at first they talked about “affine quantum mechanics (AQM)” in “conformal differential geometry” (Santamato/DeMartini 2013), then they shifted to “conformal quantum geometrodynamics (CQG)” (De Martini/Santamato 2014a,b). Physical problems for their approach, like those of negative probability densities for relativistic particles, remain outside the scope of this article.

radius of a relativistic top $a = \sqrt{6} \frac{\hbar}{mc}$, was especially important because due to it the geometry would “know” about the mass of the spinning particle. For the metric they found a constant Riemannian scalar curvature ${}_gR = \frac{6}{a^2} = \frac{(mc)^2}{\hbar^2}$ induced from the Lorentz component (De Martini/Santamato 2014a, p. 3313). This gave a surprising *Riemannian geometrization of the configuration space* of the non-quantum, relativistic top by a non-definite metric g_{ij} of signature $(3+3, 1+3)$ with constant scalar curvature.

11.4.2.3 Santamato’s Random Processes in the 1980s

With this framework established, the next stage involved analyzing particle motion in the configuration space. Santamato characterized this as a (weak) random process described by an ensemble of trajectories $q^i(t, \omega)$ with ω “the sample tag”⁶⁴ based on a well-defined and normalized probability density $\bar{\rho}(q, t)$ satisfying the continuity equation (Santamato 1984a, p. 217)

$$\partial_t \bar{\rho} + \partial_i (\bar{\rho} v^i) = 0. \tag{11.46}$$

He gave an unusual derivation for the velocity field v^i of the random process associated to a given Lagrangian $L(q, \dot{q}, t)$. After shifting the Lagrangian to $L^* = L + \frac{d}{dt}S$ for some sufficiently differentiable function $S(q, t)$,⁶⁵ he analyzed the *averaged action functional*

$$I(t_0, t_1) = E \left(\int_{t_0}^{t_1} L^*(q(t, \omega), \dot{q}(t, \omega), t) dt \right) \tag{11.47}$$

where $E(\dots)$ is the expected value. He then looked for the minimum I under variation of $v^i = \dot{q}^i$ with respect to all random motions obeying a flow equation and satisfying given initial data. As a necessary condition for the existence of such a minimum it turned out that S has to solve the *Hamilton-Jacobi* equation (Santamato 1984a, app. A)

$$\partial_t S + H(q, \nabla S, t) = 0, \tag{11.48}$$

where $H(q, p, t)$ is the classical Hamiltonian corresponding to $L(q, \dot{q}, t)$. Then the minimizing velocity field of (11.47) is the corresponding Hamilton-Jacobi flow.

Santamato could hope that the wave equations of QM might be derivable from his random processes if a suitable classical Lagrangian (should there be any) were modified in a convincing way. “Convincing” here means a change of the Lagrangian

⁶⁴That is, $\omega \in \Omega$, the sample space of a probability triple $(\Omega, \mathfrak{F}, P)$, where $\mathfrak{F} \subset \mathfrak{P}(\Omega)$ are the random events and P is a probability measure on Ω .

⁶⁵ L^* has the same Euler-Lagrange equations as the original L .

by geometrical terms, where the geometry is influenced by the particle's (random) motion. As he put it: "Geometry is not prescribed; rather it is determined by physical reality. In turn, geometry acts as a 'guidance field' for matter." (Santamato 1984a, p. 216) He argued that such a "feedback mechanism between geometry of space and particle motion" was "quite analogous" to general relativity and might lead to "a theory that is physically indistinguishable from traditional quantum mechanics" (ibid.).

At this point Santamato supplemented the Euclidean, or more generally Riemannian, structure of the configuration space by a *Weylian scale connection*. He called this a "vector transplantation law" and denoted it by ϕ_k , corresponding to our $-\varphi_k$. In the case of a non-Euclidean Riemannian component of the metric g_{ij} , the continuity equation for the adapted probability density $\rho = |g|^{-\frac{1}{2}} \bar{\rho}$ turns into the covariant equation:

$$\partial_t \rho + {}_g \nabla_i (\rho v^i) = 0 \quad (11.49)$$

A classical Lagrangian $L_c(q, \dot{q}, t)$ on the original non-relativistic configuration space could now be modified on the Weylianized space according to *Santamato's 1st postulate* (Santamato 1984a, eqn. (8)):

$$L(q, \dot{q}, t) = L_c(q, \dot{q}, t) + \gamma \frac{\hbar^2}{2m} R(q, t), \quad \text{with } \gamma = \frac{n-2}{4(n-1)}, \quad (11.50)$$

where $R(q, t)$ denotes the complete Weylian scalar curvature. With a sign-inversion convention for the scalar curvature (11.2),⁶⁶ Santamato wrote it as

$$R = {}_g R + (n-1)(n-2) \phi_i \phi^i - 2(n-1) {}_g \nabla_i \phi^i. \quad (11.51)$$

The term in R enters (11.50) like a supplement to the potential. The Hamilton-Jacobi equation of the random flow (11.48) thus becomes

$$\partial_t S + H_c(q, \nabla S, t) - \gamma \frac{\hbar^2}{m} R = 0. \quad (11.52)$$

According to the author's program, R should depend on the random process and was assumed to be time dependent. Therefore the ϕ_k cannot be arbitrarily given, but rather were to be determined by the probability density of the matter flow in the configuration space. Santamato applied his averaged least action principle (11.47) and evaluated it with (11.50) for vanishing L_c , i.e., for $R(q, t)$ alone, obtaining the encouraging result (ibid. eqn. (19))

$$\phi_i = -(n-2)^{-1} \partial_i \ln \rho. \quad (11.53)$$

⁶⁶Cf. fn 31.

The scalar curvature $R = {}_g R + {}_\varphi R$ turned out to contain (Santamato 1984a, eqn. (20))

$${}_\varphi R = \frac{1}{\gamma \sqrt{\rho}} \left({}_g \nabla_i \partial^i \sqrt{\rho} \right), \quad (11.54)$$

in striking accord with Bohm's quantum potential (11.36). Santamato jumped without hesitation from recognizing this formal agreement to a deeper realistic conclusion:

... according to Eq. [(11.53)], the geometric properties of space (...) are indeed affected by the presence of the particle itself. In turn, this alteration of geometry of space acts on the particle through the quantum force $f_i = \gamma \frac{\hbar^2}{m} \partial_i R$, which, according to Eq. [(11.51)], depends on the gauge vector and its first and second derivatives. (Santamato 1984a, 219, eqn. numbers adapted)

This was a strong statement since it suggested a close kinship between *Santamato's modification of geometry* and the *general theory of relativity* (GR), even though his modification did not refer to the spacetime manifold of GR, the "extensive medium of the world" as Weyl liked to call it, but rather to the configuration space of a dynamical system.

In his next paper Santamato derived the Klein-Gordon equation (11.38) in the same way, starting from a random process. He used a configuration space based on Minkowski space \mathbf{M} and superimposing a Weylian scale connection (11.51). Including electromagnetic terms, the Lagrangian for the *relativistic ensemble* was

$$L(x, \dot{x}) = \left(1 + \gamma \frac{\hbar^2}{(mc)^2} R(x) \right)^{\frac{1}{2}} |x| + \frac{e}{mc^2} A_\mu \dot{x}^\mu, \quad (11.55)$$

with $R = {}_\varphi R$ the Weylian scalar curvature (${}_g R = 0$). Taking into account eqn. (11.54), it followed that a complex function $\psi = \sqrt{\rho} e^{iS}$, constructed as usual (up to a factor \hbar^{-1} in the exponent) from the flow quantities, "obeys the Klein-Gordon equation" (Santamato 1984b, p. 2479). This was no small achievement; but Santamato did not continue his research along these lines for many years to come.

11.4.2.4 A Look at the Second Phase in Cooperation with De Martini

After a long interruption, Santamato, now in joint work with his colleague De Martini, gave a new derivation for the Weyl-geometric approach to the foundations of quantum mechanics. Moreover, this new series of papers gives a much clearer presentation of the underlying scale-co-/invariant structure. (De Martini/Santamato 2014a) starts from a field-theoretic Lagrangian using a metric-affine approach (cf. pp. 288ff.):

$$\mathfrak{L} = \rho (\partial_\mu \sigma \partial^\mu \sigma + \gamma \hbar^2 R) \sqrt{|g|}. \quad (11.56)$$

This involved two scalar fields ρ, σ with weights $w(\rho) = -2, w(\sigma) = 0$ under conformal rescaling and a scalar curvature term R defined with respect to a metric g_{ij} and an independently defined affine (torsion free) connection Γ_{ij}^k . The field σ now took over the role of the former Hamilton-Jacobi principal function S (De Martini/Santamato 2014a, eqn. (1)). Variation with respect to the scalar fields leads to the dynamical equations:

$$\partial_\mu \sigma \partial^\mu \sigma + \gamma \hbar^2 R = 0 \tag{11.57}$$

$$\partial_\mu \left(\sqrt{|g|} \rho \partial^\mu \sigma \right) = 0 \quad \longleftrightarrow \quad {}_g \nabla_\mu (\rho \partial^\mu \sigma) = 0. \tag{11.58}$$

Now, (11.57) has the same form as the Hamilton-Jacobi equation of an uncharged Klein-Gordon field (11.39), where the scalar curvature, up to sign, takes the place of the “quantum potential”. (11.58) may be read as a continuity equation for a flow with density ρ and velocity given by $\partial^\mu \sigma$ (if timelike). By variation with respect to the affine connection,⁶⁷ the authors concluded that the affine connection has the Weyl-geometric form (11.1) with the scale connection as in (11.53), just like the one Santamato had derived in the 1980s from his average action principle.⁶⁸

The authors did not consider a variation of the metric because they had a de Broglie–Madelung–Bohm context in mind, which meant that the Riemannian metric of the configuration space was determined by the Lagrangian of a classical system. They immediately turned to this by giving a *mechanical interpretation* of their scalar field theory. In the relativistic case, the equations (11.57), (11.58) as well as the Hamilton-Jacobi and continuity equations can be derived from a variational problem $\delta \int L d\tau = 0$ with

$$L_r = \sqrt{-\gamma \hbar^2 R(q) g_{\mu\nu} \dot{q}^\mu \dot{q}^\nu}. \tag{11.59}$$

This fit well with the program for geometrizing a configuration space with Riemannian metric as this relates to a classical process, amended by a Weylian scale connection standing in “back reaction” with a solution pair (σ, ρ) of (11.57), (11.58). With, like above, $\gamma = \frac{n-2}{4(n-1)}$ and $n = 4$, the Weylian component of the scalar curvature is in fact⁶⁹

$${}_\varphi R = \frac{1}{4\gamma} \left(2\rho^{-1} {}_g \nabla_i \partial^i \rho - \rho^{-2} \partial_i \rho \partial^i \rho \right) = \gamma^{-1} \frac{{}_g \nabla_i \partial^i \sqrt{\rho}}{\sqrt{\rho}} \tag{11.60}$$

$$= \gamma^{-1} \frac{2m}{\hbar^2} U = \gamma^{-1} Q, \tag{11.61}$$

⁶⁷Compare Section 11.6.4.1.

⁶⁸A sign error in the formula of the Weyl-geometric affine connection (De Martini/Santamato 2014a, eqn. (4)) notwithstanding.

⁶⁹(De Martini/Santamato 2014a, p. 3310)

where U and Q are the additional terms (“quantum potentials”) (11.36), (11.41) on the right hand side of the Hamilton-Jacobi equations of a Schrödinger or a Klein-Gordon particle, respectively. It has to be understood that (11.59) holds only for relativistic particles (Klein-Gordon and Dirac), while for the non-relativistic case of a Schrödinger particle with $R = {}_\varphi R$ the Lagrangian is (11.50).

For investigating *relativistic spinning* particles De Martini and Santamato considered a point dynamics with internal degrees of freedom in the configuration space $V = \mathbf{M} \times SO(3, 1)$ described in Section 11.4.2.2. The Hamilton-Jacobi equation of a process governed by the Lagrangian (11.59) plus an electromagnetic term L_{em} is given by (Santamato/DeMartini 2013, eqn. (7))

$$(\partial_\mu S - \frac{e}{c}A_\mu)(\partial^\mu S - \frac{e}{c}A^\mu) + \hbar^2\gamma R = 0, \quad (11.62)$$

where S satisfies the divergence equation

$$D_\mu(\partial^\mu S - \frac{e}{c}A^\mu) = 0 \quad (11.63)$$

with the scale-covariant derivative (here $D_\mu = \nabla_\mu - 2\phi_\mu$ in our weight convention with $w(g_{\mu\nu}) = 2$, and with ∇_μ the Weyl-geometric covariant derivative). For a current defined by $j^\mu = \chi^{-(n-2)}\sqrt{|g|}(\partial^\mu S - \frac{e}{c}A^\mu)$ this boils down to an ordinary continuity equation

$$\partial_\mu j^\mu = 0, \quad (11.64)$$

so j^μ is obviously scale invariant.

The transition to a complex wave function depending on all coordinates q of the configuration space

$$\psi(q) = \sqrt{\rho} e^{\frac{i}{\hbar}S} \quad (11.65)$$

transforms the equs. (11.62), (11.64) into the *linear differential equation* of second order

$$(\hat{p}^\mu - \frac{e}{c}A^\mu)(\hat{p}_\mu - \frac{e}{c}A_\mu)\psi + \hbar^2\gamma {}_gR\psi = 0, \quad (11.66)$$

where \hat{p} denotes the differential operator with $\hat{p}_\mu = -i\hbar\partial_\mu$. This has the form of the Klein-Gordon equation (11.38) with a mass factor which contains only the *Riemannian part* of the scalar curvature, $\hbar^2\gamma {}_gR = \hbar^2\gamma \frac{6}{a^2} = m^2c^2$. The *Weylian component* ${}_\varphi R$ is controlled via (11.53) by the density of the quantum flow. The authors commented as follows:

This is a striking result as it demonstrates that the Hamilton- Jacobi equation, applied to a general dynamical problem can be transformed into a linear eigenvalue equation, the

foremost ingredient of the formal structure of quantum mechanics and of the Hilbert space theory. (Santamato/DeMartini 2013, p. 636)

The first step towards a reconstruction of the Hilbert space quantization of the relativistic top was thus achieved.

In the next step, the authors analyzed the decomposition of a solution ψ of (11.66) into components $\psi_{u,v}$ lying in finite dimensional representations of $SO(3, 1)$ of type $D^{(u,v)}$ with $2u, 2v \in \{0, 1, 2, \dots\}$. This $\psi_{u,v}(q)$ can be factorized into functions of the spatial variable x with values in the representation space of $D^{(u,v)}$, and θ -dependent representation matrices operating on the latter. In spinor notation similar to van der Waerden’s symbolism this becomes:⁷⁰

$$\psi_{u,v}(q) = D^{(u,v)}(\Lambda(\theta))_{\sigma}^{\sigma'} \psi_{\sigma}^{\sigma'}(x) + D^{(v,u)}(\Lambda(\theta))_{\dot{\sigma}}^{\dot{\sigma}'} \psi_{\dot{\sigma}}^{\dot{\sigma}'}(x). \tag{11.67}$$

For the particular choice $u = v = \frac{1}{2}$ this leads to a pair of 2-component spinor fields on Minkowski space, equivalent to a 4-component Dirac field $\Psi(x) = \begin{pmatrix} \psi_{\sigma}^{\sigma'}(x) \\ \psi_{\dot{\sigma}}^{\dot{\sigma}'}(x) \end{pmatrix}$ in the Weyl representation. Then the equation (11.66) acquires a form which, after neglecting an extremely small term in the electromagnetic field strength,⁷¹ the authors identify as the *squared Dirac equation*

$$\mathfrak{D}_+ \mathfrak{D}_- \Psi = \mathfrak{D}_- \mathfrak{D}_+ \Psi = 0, \tag{11.68}$$

$$\text{where } \mathfrak{D}_{\pm} = \gamma^{\mu} (p_{\mu} - \frac{e}{c} A_{\mu}) \pm m$$

with the Dirac matrices γ^{μ} ($\mu = 0, \dots, 3$) (Santamato/DeMartini 2013, p. 639).⁷²

Solutions of (11.68) can be decomposed into a superposition of the linear Dirac equation with positive and negative mass. The authors proposed that the negative mass contributions have to be “disregarded as unphysical” (Santamato/DeMartini 2013, p. 641). Even without trying to assess this claim, it is clear that with their model for relativistic spinning particles Santamato and De Martini had achieved a surprising step forward for the geometrization program in line with the dBMB approach started in the 1980s. They did not stop there, however, but rather went on to investigate the nonlocality of EPR systems using this approach. Their considerations led to a justification of the spin-statistic relation usually derived by quantum field

⁷⁰For van der Waerden’s spinor symbolism see (Schneider 2011).

⁷¹This term, $\frac{e^2 q^2}{c^2} (H^2 - E^2)$, is comparable with the linear term in the field strengths only if the latter are very large, $E \sim 10^{18} \text{Vm}^{-1}$, $H \sim 10^9 \text{T}$. “To have an idea how large is this field, an electron at rest is accelerated by such field up to 10^9GeV in a linear accelerator 1 m long” (De Martini/Santamato 2014a, p. 3315).

⁷²It remains unclear to me (E.S.) how the representation matrices of the “Euler angles” of configuration space are suppressed, while the change of coordinate frames in Minkowski space gets represented on the spinor fields.

methods (De Martini/Santamato 2014a, 2015). In order not to blow this survey out of proportion, these derivations, although central for the content of their papers, have to be omitted here.

11.4.3 *An Attempt at Bridge Building to Gravity*

We have yet to review attempts to connect the dBMB approach to gravity with a specific reference to Weyl geometry. Different authors tried to do so, in particular *Fatimah Shojai*, *Ali Shojai* and *Mehdi Golshani* working in Tehran. Another, to my taste slightly more bizarre, step in this direction was taken by *Giorgio Papini* and *Robert Wood* on the occasion of a symposium honouring J.-P. Vigi er (Wood/Papini 1997). Some years earlier the latter had tried to fix a defect in Dirac’s 1972 proposal for reviving Weyl’s original interpretation of the scale connection as the electromagnetic potential, resulting from the non-integrability of the scale connection.⁷³ Papini and Wood proposed to solve this problem by considering “bubbles” in the environment of atoms that break the scale symmetry, which still holds in the large, i.e. outside the “bubble” (Wood/Papini 1992). For the Vigi er symposium they recycled their idea by establishing a connection to a dBMB approach governing the dynamics in the bubble, an idea somewhat similar to de Broglie’s proposal.

At the end of the 1990s, F. and A. Shojai, occasionally with Golshani as a third author, began investigations in which they hoped to be able to use a Bohmian approach to quantize a part of the gravitational structure (Shojai et al. 1998a,b,c; Shojai/Golshani 1998; Shojai/Shojai 2000). To do so they used methods from scalar-tensor theories of gravity together with a specific emphasis on conformal ideas that brought their work close to Weyl geometry. During a sojourn at the *Max Planck Institute for Gravitational Physics* at Potsdam they laid open this connection and proclaimed that this was the correct framework for their approach (Shojai/Shojai 2003), which we now explain.

In the 1980s *Jayant Narlikar* and *Thanu Padmanabhan* had studied a simplified version of quantum gravity that was invariant under conformal changes of the metric, $g_{\mu\nu} \mapsto \Omega^2 g_{\mu\nu}$. They proposed to *quantize only the factor Ω* , viz. the scale degree of freedom of the metric. This had the great advantage of leaving the conformal structure unaffected by the quantization, thereby circumventing the infamous obstacle of a fuzzy causal structure, which other approaches to quantum gravity encountered. On this basis Narlikar and Padmanabhan calculated semiclassical approximations for cosmological solutions of the Einstein equation (Narlikar/Padmanabhan 1983; Padmanabhan 1989).⁷⁴

⁷³Among others, this had led to the measurement problem for atomic clocks.

⁷⁴In the physics literature, and also in the paper by Shojai and Golshani, Ω is often referred to as the “conformal” degree of freedom of the metric, or even the “conformal structure”. The latter

In one of their early joint papers, F. Shojai and M. Golshani took this idea up, but in contrast to Narlikar and Padmanabhan they attempted a Bohmian path towards quantizing the scale factor (Shojai/Golshani 1998). This was quite daring because Bohmian quantum mechanics had been developed only for systems with finite degrees of freedom. Shojai and Golshani, however, invoked the idea of de Broglie to re-interpret the “quantum mass” $\mathfrak{M}_o^2 = m^2 + \frac{\hbar^2}{c^2} Q$ of a Klein-Gordon system (11.39) as a conformal modification of the Minkowski metric using $\Omega^2 = \frac{\mathfrak{M}_o^2}{m^2} = 1 + \frac{\hbar^2}{m^2 c^2} \frac{\nabla_\mu \partial^\mu \sqrt{\rho}}{\sqrt{a}}$. They considered this rescaling factor as a representative for the quantum degrees of freedom of a globally defined Klein-Gordon field.

Another problem was that \mathfrak{M}_o^2 could become negative. The Shojais and Golshani solved it by passing over to the exponential (Shojai/Golshani 1998, eqn. (12))

$$\mathfrak{M}^2 = m^2 e^{\frac{\hbar^2}{m^2 c^2} \frac{\nabla_\mu \partial^\mu \sqrt{\rho}}{\sqrt{a}}} . \quad (11.69)$$

The linear approximation coincides with \mathfrak{M}_o^2 , and the conformal factor became

$$\Omega^2 = \frac{\mathfrak{M}^2}{m^2} = e^{\frac{\hbar^2}{m^2 c^2} \frac{\nabla_\mu \partial^\mu \sqrt{\rho}}{\sqrt{\rho}}} . \quad (11.70)$$

They started from a Lagrangian with Einstein-Hilbert term and a matter Lagrangian characteristic for a classical Jacobi-Hamilton system with Hamilton-Jacobi function S and flow density ρ ,

$$\mathfrak{L}_m = \frac{\hbar^2}{m} \left(\frac{\rho}{\hbar^2} \partial_\mu S \partial^\mu S - \frac{m^2}{\hbar^2} \rho \right) \sqrt{|g|} . \quad (11.71)$$

Santamato’s viewpoint (which was apparently unknown to the Tehran authors) had been that the introduction of the quantum potential turned the corresponding dynamical system into the Hamilton-Jacobi form of a Klein-Gordon system (11.39), (11.40). The Shojais proceeded differently. Drawing support from de Broglie’s position they argued that:

... the de Broglie remark leads to the conclusion that the introduction of the quantum potential which contains the quantal behaviors of the particles is *equivalent* to the introduction of a conformal factor $\Omega^2 = \frac{\mathfrak{M}_o^2}{m^2}$ in the metric (Shojai/Golshani 1998, p. 683, emphasis E.S.).

This was a puzzling statement, though, since De Broglie had considered a geometrization for a single particle *in the absence of electromagnetic and gravitational fields* (Section 11.4.1.2). It remained unclear whether the argument could be transferred to the case of gravitational fields and in what sense such an “equivalence” was to be understood.

is clearly mistaken, whereas the first is, at best, misleading. Therefore I avoid this terminology in favour of *scaling degree of freedom*.

In Santamato's geometrization of a dBMB Hamilton-Jacobi system the "prepotential" of a Weylian scale connection on the configuration space (11.53) was $\ln \rho$, up to a constant factor. This leads to the Weylian curvature expression (11.54), which is equivalent to a Bohmian "quantum potential" in the dynamical equation (11.52). De Broglie, and with him the Shojais, used a different geometrization idea. Their "prepotential" of the Weylian scale connection was the scale factor Ω between a classical metric and the metric describing a quantum system that entered the exponential expression (11.70). Following de Broglie, one had to consider geodesic flows with the implicit constraint of initial conditions orthogonal to a level surface of a related Hamilton-Jacobi principal function S . The modification (11.69) of the usual "quantum mass" formula implies that one cannot expect equivalence in the literal sense. Even if one wants to read the argument as a motivation for a new type of dBMB-like quantization procedure, following the de Broglie paradigm, a justification for the attempted generalization from de Broglie's case (no gravitation) to the general case had to be given.

Nevertheless, these authors did not hesitate to take this step as a starting point for investigating cosmological models in which matter fields were given in different versions of scalar-tensor theories.⁷⁵ As a result, a Klein-Gordon field appeared on large scales rather than as a descriptor of the motion of a single quantum particle. In some papers it played the role of a matter field (Shojai/Golshani 1998, p. 683), (Shojai 2000, p. 1762), in others that of a "quantum gravity" modification of the metric field (Shojai et al. 1998a, p. 2728). The Shojais were convinced that "the correct quantum conformal degree of freedom would be achieved, and . . . that the theory works for a particle as well as for a real ensemble of the particle under consideration and that it includes pure quantum gravity effects" (Shojai/Shojai 2000, 1763). But it remained unclear what "quantum gravity" means here.

One of the papers written by all three authors dealt explicitly with conformal transformations in scalar-tensor theories (Shojai et al. 1998a).⁷⁶ They distinguished between a "background metric", in which quantum effects were encoded by the varying "quantum mass" \mathfrak{M} , and a "physical metric", for which \mathfrak{M} was rescaled to a constant value \bar{m} . Then "some part of the curvature of space-time represents the quantum effects" (Shojai et al. 1998a, p. 2726). Leaving aside this physical interpretation and the reasonability of this approach, one might wonder whether a reformulation in terms of Weyl geometry might at least help to clarify the mathematical import of such statements.

This is what A. and F. Shojai attempted in (Shojai/Shojai 2003) and in a preprint that followed (Shojai/Shojai 2004). In the meantime they had adopted Dirac's theory of 1972 (see Section 11.3.2.1), but did not follow Dirac's *em* dogma. Rather they considered the scale connection as "a part of the geometry of the space-time",

⁷⁵(Shojai/Golshani 1998; Shojai et al. 1998a,b; Shojai/Shojai 2000, 2001).

⁷⁶The authors differentiated between "scale transformations" and "conformal transformations". In their terminology the first operated only on the metric, while the latter rescaled all physical fields according to their weights.

implicitly constrained in their context by the integrability condition.⁷⁷ But without much hesitation they declared that Dirac's scalar field β in (11.15) "represents the quantum mass field" in the sense of their embryonic theory outlined above (Shojai/Shojai 2003, p. 7 preprint). They did so, however, without discussing how the different Lagrangians for the Dirac field and their Klein-Gordon fields could be related to one other. Only a rather opaque perturbative argument was given as to why a solution of the β -scalar field equation can be identified with an expression of the "quantum mass" type, $\beta \mapsto \mathfrak{M}$ (Shojai/Shojai 2003, p. 13f).⁷⁸ On the other hand, this identification at least somewhat clarified their discussion of different frames. They now considered "different conformal frames" as "identical pictures of the gravitational and quantum phenomena" (ibid., p. 9).⁷⁹

In light of such open issues, A. and F. Shojai's conclusion that Weyl geometry "provides a unified geometrical framework for understanding the gravitational and quantum forces" (Shojai/Shojai 2003, p. 10 preprint) was at least premature and reads like a rather grand speculation. Not all readers had this impression, however, and their program found at least one active successor, *R. Carroll* (Carroll 2004).⁸⁰ But the critical points required for a justification of the "Tehran program" seem not to be clarified in this work either.

11.5 Scale Covariance in the Standard Model of Elementary Particle Physics

About the middle of the 1970s the standard model of elementary particle physics (SM) started to become widely accepted as the key to the basic structures of matter (Kragh 1999, chap. 22), (Pickering 1988). Besides the point-dependent (localized) *internal* symmetries of the electroweak forces, $SU(2) \times U(1)$, and the chromodynamic symmetry of the strong forces $SU(3)$, the new paradigm of gauge field theories worked with the non-localized ("global") *external* symmetries of special relativity, the Lorentz group. Characteristic for the paradigm was a global, though only approximately respected scale invariance of the field Lagrangians, broken by the mass term of the Higgs field alone. The Higgs field Φ , a scalar field with values in an isospin $\frac{1}{2}$ representation of the electroweak group, was the clue for making electroweak symmetry of elementary particles consistent with mass terms. The latter was understood as a "spontaneous breaking" of the electroweak symmetry

⁷⁷The Dirac Lagrangian was stripped of the Yang-Mills term of the scale connection (Shojai/Shojai 2004, eqn. 6).

⁷⁸The claim that one can identify a Dirac-type scalar field with a "quantum mass" field remains, in my view, an unfounded speculation; E.S.

⁷⁹The authors even conventionalized this idea as a "conformal equivalence principle" (Shojai/Shojai 2003, p. 10), (Shojai/Shojai 2004, p. 63).

⁸⁰Not "S. Carroll", as listed on occasion in the bibliography of later papers.

and came to be known as the “Higgs mechanism” (Borrelli 2015). In this section we look at some attempts to bridge the gap between the Higgs field and the scalar field of gravity.

11.5.1 Englert, Smolin and Cheng, 1970/80s

11.5.1.1 A Conformal Approach

One of the originators of this theory of the Higgs mechanism, *Franç Englert*,⁸¹ tried to play a similar game with “spontaneous symmetry breaking” in gravity, here using a real-valued scalar field with scale symmetry in the sense of conformal rescaling. In a paper written together with *Edgar Gunzig*, *C. Truffin* and *P. Windey*, the authors established an explicit link to JBD gravity (Englert et al. 1975). But in contrast to (Deser 1970), Englert and his coworkers considered conformal gravity as part of the quantum field program. They assumed a “dimensionless”, i.e. scale invariant, Lagrangian for gravitation with a square curvature term derived from an affine connection Γ but *not* bound to the metric, $\mathcal{L}_{\text{grav}} = R^2 \sqrt{|\det g|}$, in addition to a Lagrangian matter term (Englert et al. 1975). As a consequence, they varied with respect to the metric g and the connection Γ independently.

“To make contact with General Relativity” (p. 74), the authors assumed that the scalar curvature was expressed by a scalar function $R \sim \omega^2$ (they used the symbol φ instead of ω). The Euler-Lagrange equation of the affine connection resulted in a relation like (11.1) for the Weyl-geometric case, with an integrable scale connection $\varphi = d \log \omega$ (Englert et al. 1975, eqs.(7), (8)). By such a specialization, this approach looked much like a Weyl structure, but this was not the point of view of the authors. They proceeded as much as possible on a purely “conformal” line in their search for connecting paths between quantum field theory of scalar fields and general relativity. After some tentative quantum considerations, they came back to a “classical phenomenological description” of their theory (Englert et al. 1975, 76). For this description they introduced a scalar field $\phi(x) = \lambda^{-1} e^{\lambda \sigma(x)}$ coupled to gravity as in eqn. (11.12), with the necessary specification $\xi = \frac{1}{6}$ in order to secure conformal symmetry (Englert et al. 1975, eqn. (16)). They considered σ to be a “dilation field” (sic!) representing a “Nambu-Goldstone boson” coupling to the mass terms. After some twists and turns, they summed up by asserting that their original action principle

... matches all the results of General Relativity at a classical level, provided mass originates in dynamical breakdown of symmetry. Thus, the fundamental finite component fields must be massless and of the kind currently used in gauge field theories, but without scalar mesons (Englert et al. 1975, 76).

⁸¹See, e.g. Karaca (2013).

In a follow-up paper with Truffin, Englert studied the perturbative behaviour of his version of conformal gravity ($\xi = \frac{n-2}{4(n-1)}$) coupled to massless fermions and photons in $n \geq 4$ dimensions.⁸² The two authors came to the conclusion that by using their approach anomalies arising in the calculations using non-conformal actions disappeared at the tree and 1-loop levels. They took this as an indicator that gravitation might perhaps arise in a “natural way from spontaneous breakdown of conformal invariance” (Englert et al. 1976, 426).

11.5.1.2 Smolin Introduces Weyl Geometry

The paper (Englert et al. 1976) was one of the early steps in the direction (i) of our introduction. Other authors followed and extended this view, some working explicitly in a Weyl-geometric setting, others continuing to use the language of conformal geometry. The first strategy was chosen by *Lee Smolin* in his paper (Smolin 1979). In section 2 of that paper he gave an explicit and clear introduction to Weyl geometry.⁸³ What he called “conformally metric gravitation” was built on a matter-free Lagrangian using Weyl-geometric curvature terms R , $Ric = (R_{\mu\nu})$, $f = (f_{\mu\nu})$ for scale curvature and a gravitational Lagrangian of order two. With a slight adaptation of notation by way of the scale-covariant Weylian derivatives D , this was (Smolin 1979, eqn. (13)):

$$\begin{aligned} |det g|^{-\frac{1}{2}} \mathcal{L}_{\text{grav}} = & -\frac{1}{2} c \phi^2 R + [-e_1 R^{\mu\nu} R_{\mu\nu} - e_2 R^2] \\ & + \frac{1}{2} D^\mu \phi D_\mu \phi - \frac{1}{4g^2} f_{\mu\nu} f^{\mu\nu} - \lambda \phi^4, \end{aligned} \quad (11.72)$$

where c, e_1, e_2, g, λ are coupling coefficients.⁸⁴ For coefficients of the quadratic curvature terms (in square brackets) with $e_2 = -\frac{1}{3}e_1$, the latter is variationally equivalent (equal up to divergence) to the squared conformal curvature $C^2 = C_{\mu\nu\kappa\lambda} C^{\mu\nu\kappa\lambda}$.⁸⁵

Smolin introduced the scalar field ϕ not only for formal reasons (“to write a conformally invariant Lagrangian with the required properties”), but also with a

⁸²The motivation for considering $n \geq 4$ was the method of dimensional regularization for the quantization of the theory.

⁸³In his bibliography he went back directly to (Weyl 1922) and (Weyl 1918a); he did not quote any of the later literature on Weyl geometry.

⁸⁴Signs have to be taken with caution. They may depend on conventions for defining the Riemann curvature, the Ricci contraction, and the signature. Smolin, e.g., used a different sign convention for *Riem* to the one used in this survey. Signs given here are adapted to *signature* $g = (3, 1)$. The Riemann tensor and Ricci contraction are those usually adopted in the mathematical literature, see fn. 31.

⁸⁵This seems to have been widely known. For an explicit statement see, e.g., (Hehl et al. 1996).

physical interpretation in mind similar to that given in (Englert et al. 1976),⁸⁶ namely “as an order parameter to indicate the spontaneous breaking of the conformal invariance” (Smolin 1979, 260). His Lagrangian used a modified adaptation from JBD theory “with some additional couplings” between the scale connection φ and the scalar field ϕ . Smolin emphasized that “these additional couplings go against the spirit of Brans-Dicke theory” because from the Riemannian point of view they introduced a non-vanishing divergence of the non-gravitational fields. For low energy considerations, Smolin dropped the square curvature term (square brackets in (11.72)), adding an “effective” potential term of the scalar field $V_{\text{eff}}(\phi)$ to derive the equations of motion by varying with respect to g , ϕ , and φ . He was also able to obtain the Einstein equation, scalar field equation, and Yang-Mills equation for the scale connection.

Smolin’s reduced Lagrangian contained terms in the scale connection:⁸⁷

$$-\frac{1}{4g^2} f_{\mu\nu} f^{\mu\nu} + \frac{1}{8}(1+6c)F^2 \varphi_\mu \varphi^\mu, \quad (11.73)$$

where F denotes the constant value of ϕ in the scalar-field gauge. That looked like a mass term for the scale connection φ , the potential of the scale curvature field $f_{\mu\nu}$, which Smolin called the “Weyl field”. By comparison with the Lagrangian of the Proca equation in electromagnetic theory, Smolin concluded that the “Weyl field” has mass close to the Planck scale, given by

$$M_\varphi^2 = \frac{1}{4}(1+6c)F^2. \quad (11.74)$$

He commented that in his Weyl-geometric gravitation theory “general relativity couples to a massive vector field” φ . The scalar field ϕ , on the other hand, “may be absorbed into the scalar parts” of $g_{\mu\nu}$ and φ_μ ⁸⁸ by a change of variables and “remains massless” (Smolin 1979, 263). In this way, Smolin brought Weyl geometric gravity closer to the field theoretic frame of particle physics. He did not discuss mass and interaction fields of the SM; moreover, the huge mass of the “Weyl field” must have appeared a real nuisance.

⁸⁶(Englert et al. 1975) was not quoted by Smolin.

⁸⁷In scalar-field gauge with $\phi \doteq \phi_o = F$, his reduced Lagrangian (square gravitational terms dropped) was (Smolin 1979, eqn. (3.17))

$$|\det g|^{-\frac{1}{2}} \mathcal{L}_{\text{grav}} \doteq -\frac{1}{2}c F^2 {}_gR - \frac{1}{4g^2} f_{\mu\nu} f^{\mu\nu} + \frac{1}{8}(1+6c)F^2 \varphi_\mu \varphi^\mu - V_{\text{eff}}(F).$$

⁸⁸It is possible to choose the scale gauge such that ϕ becomes constant (scalar-field gauge, see Section 11.2.1).

11.5.1.3 Interlude

At the time Smolin's paper appeared the program of so-called *induced gravity* entered an active phase. Its central goal was to derive the action of conventional or modified Einstein gravity from an extended scheme of standard model type quantization. Among the authors involved in this program, *Stephen Adler* and *Anthony Zee* stand out, but we cannot go into this story here.⁸⁹

Smolin's view that the structure of Weyl geometry might be suitable for bringing classical gravity into a coherent alignment with standard model physics did not find much response, but it was "rediscovered" at least twice (along with an independently developed conformal version). In 1987/88 Hung Cheng at MIT, and a decade later Wolfgang Drechsler and Hanno Tann in Munich, arrived at similar insights and established an explicit extension of Weyl-geometric gravity to standard model (SM) fields (Cheng 1988; Drechsler/Tann 1999; Drechsler 1999). Simultaneous with Cheng, Moshé Flato (Dijon) and Ryszard Rącka (then at Trieste) discovered the core idea once more, although they formulated it in a strictly conformal framework without Weyl structure (Flato/Rącka 1988). None of these authors seem to have known Smolin's proposal (at least none of them cited him), nor did they refer to the papers of the others.⁹⁰ All three approaches had their own distinctive merits. Here we can give only give a short presentation of the main points of these works directly related to Weyl geometry.

11.5.1.4 Hung Cheng and His "Vector Meson"

Hung Cheng started from reading a paper by Chen Ning Yang on Einstein in which his debate with Weyl was mentioned in passing (Yang 1980). Without delving into the earlier literature, he decided to reconstruct the scale-invariant theory on his own and extended it to the electroweak sector of the SM.⁹¹ In place of Utiyama's complex scalar field ϕ (which he did not know of) he used the *Higgs field* Φ , again of weight -1 but now with values in an isospin $\frac{1}{2}$ representation, and coupled it to the Weyl geometric scalar curvature R . For the Lagrangians he postulated:⁹²

⁸⁹ For a survey of the status of investigations in 1981 see (Adler 1982); but note in particular (Zee 1982, 1983). In fact, Zee's first publication on the subject preceded Smolin's. (Zee 1979) was submitted in December 1978 and published in February 1979; (Smolin 1979) was submitted in June 1979. The topic of "origin of spontaneous symmetry breaking" by radiative correction was much older (Borrelli 2015; Karaca 2013). A famous paper was (Coleman/Weinberg 1973).

⁹⁰ Flato/Rącka's paper appeared as a preprint of the *Scuola Internazionale Superiore di Studi Avanzati*, Trieste, in 1987; the paper itself was submitted in December 1987 to *Physics Letters B* and published in July 1988. Cheng's paper was submitted in February 1988, published in November. Only a decade later, in March 2009, Drechsler and Tann got acquainted with the other two papers. This indicates that the Weyl-geometric approach in field theory had not yet acquired the coherence of a research program with a stable communication network.

⁹¹ Personal communication to ES, April 04, 2017.

⁹² In the sequel the isospin extended scalar field will be denoted by Φ .

$$\mathcal{L}_R = \frac{1}{2} \beta \Phi^* \Phi R |det g|^{\frac{1}{2}} \quad (11.75)$$

$$\mathcal{L}_\Phi = \frac{1}{2} \tilde{D}^\mu \Phi^* \tilde{D}_\mu \Phi |det g|^{\frac{1}{2}}. \quad (11.76)$$

The scale-covariant derivatives were extended to a localized electroweak (*ew*) group $SU(2) \times U(1)$. With the usual notations from the standard model, W_μ^j the field components of the $su(2)$ part (with respect to the Pauli matrices σ_j ($j = 0, 1, 2$)) and B_μ for $u(1)_Y \cong \mathbf{R}$, and with coupling coefficients g, g' the derivative reads⁹³

$$\tilde{D}_\mu \Phi = (\partial_\mu - \varphi_\mu + \frac{1}{2} ig W_\mu^j \sigma_j + \frac{1}{2} g' B_\mu) \Phi. \quad (11.77)$$

The sign of the kinetic term of the Higgs field (11.76) shows that Cheng supposed *sig* $g = (+ - --)$, which agrees with the high energy context, while the sign of (11.75) indicates that he used the sign-inverted convention for curvature.⁹⁴ He added Yang-Mills interaction Lagrangians for the *ew* interaction fields F and G of the potentials W (values in su_2), respectively B (values in $u(1)_Y$), and added a scalar curvature term in $f = (f_{\mu\nu}) = d\varphi$ to obtain:

$$\mathcal{L}_{\text{YM}} = -\frac{1}{4} (f_{\mu\nu} f^{\mu\nu} + F_{\mu\nu} F^{\mu\nu} + G_{\mu\nu} G^{\mu\nu}) |det g|^{\frac{1}{2}}. \quad (11.78)$$

Finally, he introduced spin $\frac{1}{2}$ fermion fields ψ with the weight convention $w(\psi) = -\frac{3}{2}$ and a Lagrangian \mathcal{L}_ψ similar to the one formulated later by Drechsler, discussed below (11.82).⁹⁵

Thus Cheng's general relativistic scalar field Φ resembled very much the Higgs field of the SM, which at that time was still a highly hypothetical object. He called the scale connection, respectively its curvature, the *Weyl meson field*. With great surprise he noticed that the scale connection does not influence the equation of motion of the spinor fields and concluded that "Weyl's vector meson does not interact with leptons or quarks. Neither does it interact with other vector mesons. The only interaction the Weyl's meson has is that with the graviton"⁹⁶ (Cheng 1988, 2183).

Because of the tremendously high mass of "Weyl's vector meson" Cheng conjectured that even such a minute coupling might be of some cosmological import. More precisely, he wondered "whether Weyl's meson may account for at

⁹³Cheng added another coupling coefficient for the scale connection, which is here suppressed.

⁹⁴See footnote 31.

⁹⁵The second term in (11.82) is missing in Cheng's publication. That is probably not intended, but a misprint. Moreover he did not discuss scale weights for Dirac matrices in the tetrad approach.

⁹⁶Remember that the φ terms of scale-covariant derivatives in the Lagrangian of spinor fields cancel.

least part of the dark matter of the universe” (ibid.), but did not pursue this idea further. At that time dark matter was just beginning to be taken seriously. But during the next decades the awareness grew and similar conjectures were stated again and again whenever theoretical entities were encountered that might represent massive particles for which experimental evidence was lacking.

11.5.1.5 Can Gravity Do What the Higgs Does?

In the same year in which Cheng’s paper appeared, *Moshé Flato* and *Ryszard Rączka* sketched an approach in which they put gravity into a quantum physical perspective.⁹⁷ In our context, the significance of this paper comes from the fact that it introduced a scale-covariant Brans-Dicke-like field in an isospin representation similar to Hung Cheng’s, but in a strictly conformal framework (Flato/Rączka 1988).

Six years later, R. Rączka took up this thread again, now in cooperation with *Marek Pawłowski*. In the meantime Pawłowski had joined the research program with a paper in which he addressed the question as to whether gravity “can do what the Higgs does” (Pawłowski 1990). In a couple of preprints⁹⁸ and two refereed papers (Pawłowski/Rączka 1994b, 1995b) these two physicists proposed a “Higgs free model for fundamental interactions”, as they described it. This proposal was formulated in a strictly conformal setting. Although it seems interesting from a theoretical point of view, we cannot discuss it here in more detail, and with the experimental detection of a Higgs scalar particle this approach has lost much of its appeal.

11.5.2 Mass Generation and Weyl-Geometric Gravity “in Munich”, 1980/90s

11.5.2.1 1990: Drechsler and Tann

A view closer to Cheng’s for establishing a connection between gravity and electroweak fields in the framework of Weyl geometry was developed a decade later by *Wolfgang Drechsler* and his PhD student *Hanno Tann* in Munich. Drechsler had been active for more than twenty years in differential geometric aspects of field theory.⁹⁹ In cooperation with *D. Hartley*, he developed an approach of his own to Weyl-geometric gravity evolving from investigations in Kaluza-Klein theories (Drechsler/Hartley 1994). Tann, coming from a background interest in geometric properties of the de Broglie-Bohm interpretation of quantum mechanics

⁹⁷More than a decade earlier Flato had sketched a covariant (“curved space”) generalization of the Wightman axioms (Flato/Simon 1972), different from the one discussed by R. Wald in this volume.

⁹⁸(Pawłowski/Rączka 1994a, 1995a,b)

⁹⁹For example (Drechsler/Mayer 1977).

(see Section 11.4.1), joined the activity a little later during work on his PhD thesis (Tann 1998). In their joint work (Drechsler/Tann 1999) as well as in their separate publications (Tann 1998; Drechsler 1999) Weyl-geometric structures were used in a coherent way, clearer than in most of the other physical papers discussed up to now.

Whereas Tann studied a complex-valued scalar field Φ , Drechsler (also in his joint paper with Tann) investigated a scalar field with values in an isospin $\frac{1}{2}$ representation of the ew group with gravitational Lagrangian

$$\mathcal{L}_{\text{grav}} = \mathcal{L}_R + \mathcal{L}_{R^2}, \quad (11.79)$$

where $\mathcal{L}_{R^2} = \tilde{\alpha} R^2 \sqrt{|det g|}$ and $\mathcal{L}_R = \frac{1}{12} \Phi^* \Phi R$ (Drechsler 1999). A common form of their linear gravitational Lagrangian, with modified Hilbert term \mathcal{L}_R and the kinetic term of the scalar field, is

$$\mathcal{L}_{R,\Phi} = \frac{\beta}{2} \Phi^* \Phi R + \frac{1}{2} (D_\nu \Phi)^* D^\nu \Phi, \quad \beta = \frac{1}{6}. \quad (11.80)$$

Here Φ^* is the adjoint (often written as Φ^\dagger), which in the case of Tann reduces to complex conjugation (often $\bar{\Phi}$), R is the Weyl-geometric scalar curvature, signature of g $(1, 3) \sim (+ - - -)$ and D_ν is the scale-covariant derivation, in Drechsler's case extended to the electroweak bundle.¹⁰⁰ In such a Lagrangian they tried to straddle the gap between the gravitational scalar field and a Higgs-like scalar field of electroweak theory.

Both authors arrived at a scale-covariant expression for the (metrical) energy momentum tensor of the scalar field ($w(T) = -2$) including terms, here with factors β^{-1} , which result from varying the scale-invariant Hilbert-Einstein term:¹⁰¹

$$T_\Phi = D_{(\mu} \Phi^* D_{\nu)} \Phi - \beta^{-1} D_{(\mu} D_{\nu)} |\Phi|^2 \quad (11.81)$$

$$- g_{\mu\nu} \left(\frac{1}{2} D^\lambda \Phi^* D_\lambda \Phi - \beta^{-1} D^\lambda D_\lambda (\Phi^* \Phi) + V(\Phi) \right).$$

Drechsler noticed that the β^{-1} terms are identical to those introduced in (Callan et al. 1970) for “improving” the “energy-momentum tensor” of a scalar field by quantum physical considerations.

In their joint paper, Drechsler and Tann introduced fermionic Dirac fields into the analysis of Weyl geometry (Drechsler/Tann 1999). Their gravitational Lagrangian

¹⁰⁰(Tann 1998, eqn. (372)), (Drechsler 1999, eqn. (2.29)). Both authors used coefficients as in the case of conformal coupling in Riemannian geometry, $\beta = \frac{1}{6}$. In the Weyl-geometric framework this was an unnecessary restriction but it suppressed the mass factor at the Planck scale for the Weyl field φ . Tann wrote the modified Hilbert term with a negative sign, because he used the sign inverted convention for the Riemann tensor, see footnote 31.

¹⁰¹ (Tann 1998, eqn. (372)), (Drechsler 1999, eqn. (2.46)).

had the form (11.79).¹⁰² For the development of a Weyl-geometric theory of the Dirac field, they introduced an adapted Lagrangian

$$\mathcal{L}_\psi = \frac{i}{2} (\psi^* \gamma^\mu D_\mu \psi - D_\mu^* \psi^* \gamma^\mu \psi) + \gamma |\Phi| \psi^* \psi \quad (11.82)$$

with (scale-invariant) coupling constant γ and Dirac matrices γ^μ with symmetric product $\frac{1}{2}\{\gamma^\mu, \gamma^\nu\} = g^{\mu\nu} \mathbf{1}$ (Drechsler/Tann 1999, (3.8)). Here the covariant derivative had to be lifted to the spinor bundle, It included an $U(1)$ electromagnetic potential $A = (A_\mu)$,

$$D_\mu \psi = \left(\partial_\mu + i \tilde{\Gamma}_\mu + \frac{iq}{\hbar c} A_\mu \right) \psi, \quad (11.83)$$

where q is the electric charge of the fermion field, and $w(\psi) = -\frac{3}{2}$, $\tilde{\Gamma}$ the spin connection lifted from the Weylian affine connection.¹⁰³ This amounted to a (local) construction of a spin $\frac{1}{2}$ bundle. Assuming the underlying spacetime M to satisfy the conditions of a spin manifold, they worked in a Dirac spin bundle \mathcal{D} over the Weylian manifold $(M, [(g, \varphi)])$. Its structure group was $G = Spin(3, 1) \times R^+ \times U(1) \cong Spin(3, 1) \times \mathbb{C}^*$, where $\mathbb{C}^* = \mathbb{C} \setminus 0$.¹⁰⁴

The two authors considered (11.82) as the Lagrangian of a “massless” theory, because the masslike factor of the spinor field $\gamma|\Phi|$ was scale invariant.¹⁰⁵ They proposed to proceed to a theory with masses by introducing a “scale symmetry breaking” Lagrange term

$$\mathcal{L}_B \sim \frac{R}{6} + \left(\frac{mc^2}{\hbar} \right)^2 |\Phi|^2 \quad (11.84)$$

¹⁰²In the appendix Drechsler and Tann showed that the squared Weyl-geometric conformal curvature $C^2 = C_{\lambda\mu\nu\rho} C^{\lambda\mu\nu\rho}$ arises from the conformal curvature of the Riemannian component ${}_g C^2$ by adding a scale curvature term: $C^2 = {}_g C^2 + \frac{3}{2} f_{\mu\nu} f^{\mu\nu}$ (Drechsler/Tann 1999, (A 54)). So one may wonder, why they did not replace the square term \mathcal{L}_{R^2} by the Weyl-geometric conformal curvature term $\mathcal{L}_{\text{conf}} = \tilde{\alpha} C^2 \sqrt{|det g|}$.

¹⁰³(11.82) can equivalently be written with a Weylianized scale-covariant derivative $\bar{D}_\mu = \left(\partial_\mu + i \tilde{\Gamma}_\mu + w(\psi) \varphi_\mu + \frac{iq}{\hbar c} A_\mu \right)$. Because φ_μ is real, the scale connection terms $w(\psi) \varphi_\mu$ in the Lagrangian cancel.

¹⁰⁴One could then just as well consider a complex valued connection $z = (z_\mu)$ with values $z_\mu = \varphi_\mu + \frac{i}{\hbar c} A_\mu$ in $\mathbb{C} = \mathfrak{L}ic(\mathbb{C}^*)$ and weight $W(\psi) = (-\frac{3}{2}, q)$. Then $D_\mu \psi = (\partial_\mu + \tilde{\Gamma}_\mu + W(\psi) z_\mu) \psi$, presupposing an obvious convention for applying $W(\psi) z$.

¹⁰⁵This argument is possible, but not compelling: $\gamma|\Phi|$ has the correct scaling weight of mass and may be considered as such.

with fixed (non-scaling) m (Drechsler/Tann 1999, sec. 4).¹⁰⁶ But they did not associate such a transition from a (seemingly) “massless” theory to a massive one by appealing to some kind of hypothetical “phase transition”. At the end of the paper they even commented:

It is clear from the role the modulus of the scalar field plays in this theory (...) that the scalar field with nonlinear selfcoupling is not a true matter field describing scalar particles. It is a universal field necessary to establish a scale of length in a theory and should probably not be interpreted as a field having a particle interpretation. (Drechsler/Tann 1999, 1050)

Their interpretation of the scalar field Φ was thus more geometric rather than viewing it as an ordinary quantum field; but their term (11.84) was introduced quite ad-hoc.¹⁰⁷

11.5.2.2 Drechsler on Mass Acquisition of Electroweak Bosons

Shortly after publication of his joint article with Tann, Drechsler extended the investigation to a gravitationally coupled electroweak theory (Drechsler 1999). Covariant derivatives were lifted as \tilde{D} to the electroweak bundle. This included the additional connection components and coupling coefficients g and g' with respect to $SU(2)$ and $U(1)_Y$, just like in Cheng’s work (11.77). The Weyl-geometric Lagrangian could now be generalized and transferred to the electroweak bundle (Drechsler 1999, (2.29)),

$$\mathcal{L} = \mathcal{L}_{\text{grav}} + \mathcal{L}_{\phi} + \mathcal{L}_{\psi} + \mathcal{L}_{\text{YM}}, \quad (11.85)$$

with contributions like those in (11.79), (11.76), (11.82), and (11.78) (*ew* terms only). Lagrangians for the fermion fields had to be rewritten so that their form was similar to the case with electromagnetic Dirac fields (11.82); they could then be decomposed into their chiral left and right contributions.

In principle, Drechsler’s proposal coincided with Cheng’s; but he proceeded with more care and by giving detailed and explicit constructions. He derived the equations of motion with respect to all dynamical variables (Drechsler 1999, eqs. (2.35) – (2.41)) and calculated the energy-momentum tensors of all fields occurring in the Lagrangian. The symmetry reduction from the electroweak group G_{ew} to the electromagnetic $U(1)_{\text{em}}$ could then be expressed by a procedure similar to that used for the standard model. $SU(2)$ gauge freedom allows one to chose a (local) trivialization of the electroweak bundle such that the Φ field assumes the form considered in the ordinary Higgs mechanism

$$\hat{\Phi} \doteq \begin{pmatrix} 0 \\ \phi_o \end{pmatrix}, \quad (11.86)$$

¹⁰⁶A similar approach is already used in Tann’s PhD dissertation.

¹⁰⁷Note that one could just as well do without (11.84) and proceed with fully scale-covariant masses – compare last footnote.

where Φ_o denotes a real valued field, and “ \doteq ” denotes the equality in a specific gauge. $\hat{\Phi}$ has the isotropy group $U(1)$ considered as $U(1)_{em}$ and was called the *electromagnetic gauge* of Φ .¹⁰⁸

In two respects Drechsler went beyond what had been done before. He *reconsidered the standard interpretation* of symmetry breaking by the Higgs mechanism (Drechsler 1999, 1345ff). And he calculated the consequences of nonvanishing *electroweak curvature components* for the *energy-momentum tensor* of the scalar field $\hat{\Phi}$ (Drechsler 1999, 1353ff). With regard to the first point, he made clear that he saw nothing compelling in the interpretation of symmetry reduction as “spontaneous symmetry breaking due to a nonvanishing vacuum expectation value of the scalar field” (Drechsler 1999, 1345). After analyzing this situation, he came to the conclusion that the transition from Φ to $\hat{\Phi}$ can be regarded as a “choice of coordinates” for the representation of the scalar field in the theory. This has, in the first place, nothing to do with a “vacuum expectation value” of this field,¹⁰⁹ since “this choice is actually not a breaking of the original \tilde{G} gauge symmetry [our G_{ew} , E.S.] but a different realization of it” (ibid.). He compared the stabilizer $U(1)_{em}$ of $\hat{\Phi}$ with the “Wigner rotations” in the study of the representations of the Poincaré group. With regard to the second point, the energy-momentum tensor of the scalar field could be calculated roughly the same way as in the simpler case of a complex scalar field, (11.81). This differed from the pseudo-Riemannian case, since the covariant derivatives $D_\mu\Phi$ etc. in (11.81) were dependent on scale or $U(1)_{em}$ curvature.

After breaking the Weyl symmetry by a Lagrangian of the form (11.84) (ibid. sec. 3), Drechsler calculated the curvature contributions induced by the Yang-Mills potentials of the *ew* group, from which he drew consequences for the energy-momentum tensor T_Φ of the scalar field. Typical contributions to components of T_Φ had the form of mass terms

$$m_W^2 W_\mu^{+*} W^{-\mu}, \quad m_Z Z_\mu^* Z^\mu, \quad \text{with} \quad m_W^2 = \frac{1}{4} g^2 |\Phi_o|^2, \quad m_Z^2 = \frac{1}{4} g_o^2 |\Phi_o|^2, \tag{11.87}$$

$g_o^2 = g^2 + g'^2$, for the bosonic fields W^\pm, Z corresponding to the generators τ_\pm, τ_o of the electroweak group (Drechsler 1999, 1353ff).¹¹⁰ These are identical with the mass expressions for the W and Z bosons in conventional electroweak theory. According to Drechsler, the terms (11.87) in T_Φ indicate that the “boson and fermion mass terms appear in the total energy-momentum tensor” through the energy tensor of the scalar field after “breaking the Weyl symmetry”.¹¹¹

¹⁰⁸In the literature it is often also called “unitary gauge”.

¹⁰⁹Mathematically speaking, it amounts to a change of trivialization of the $SU(2) \times U(1)$ -bundle.

¹¹⁰ $W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp iW_\mu^2), Z_\mu = \cos \Theta W_\mu^3 - \sin \Theta B_\mu$.

¹¹¹One has to be careful, however. Things become more complicated if one considers the trace. In fact, $tr T_\Phi$ contains a mass terms of the Dirac field of form $\gamma|\Phi_o|\hat{\psi}^*\hat{\psi}$, with γ coupling constant of the Yukawa term ($\hat{\psi}$ indicating electromagnetic gauge). One of the obstacles for making quantum matter fields compatible with classical gravity is the vanishing of $tr T_\psi$, in contrast to the (nonvanishing) trace of the energy momentum tensor of classical matter. Might Drechsler’s

Drechsler and Tann studied their scalar fields (complex or Higgs-like) as possibilities for an extension of the gravitational structure of spacetime. In their scale-covariant theory of mass acquisition they tried to understand how *mass generation* is linked to the gravitational structure. Drechsler added that in his view the scalar field "...should probably not be interpreted as a field having a particle interpretation" (Drechsler/Tann 1999, 1050). This was an interesting remark at a time when elementary particle physicists started to collect information on a possible scalar boson of the Higgs field. But the empirical confirmation of the existence of a Higgs-like boson was still far from sight; it did not materialize before the LHC started to operate at a sufficiently high level of energy and luminosity in 2012.¹¹² Even so, a more indirect link between the Higgs field and gravity, but compatible with a bosonic interpretation of the scalar field, would remain an interesting alternative perspective to that of Drechsler and Tann. Back in the 1990s, they did not anticipate that the search for a bosonic quantum of Higgs type would ever be confirmed by experiment.

11.5.3 *The "Higgs" and Weyl Scaling After 2000*

In the years following the onset of the new millennium, but still before a Higgs-like boson would be observed at the LHC, different authors continued to explore the near-to-scale invariance of the standard model. Several attempted to bridge the gap between the SM and gravity, keeping as closely as possible to the original Higgs "mechanism" developed in the special relativistic framework. These researchers did not adhere to a common research program; they used different geometric/conceptual frameworks and often took differing perspectives. Weyl-geometric methods did not always stand in the center of these investigations; some scientists worked with global scale invariance and unimodular gravity (Shaposhnikov/Zenhäusern 2009b,a), others preferred a conformal approach without making use of Weyl geometric concepts (Meissner 2009; Bars et al. 2014), and some started from conformal symmetry but were mainly interested in models with radiative breaking of scale symmetry (Foot et al. 2007a,b; Foot/Kobakhidze 2013). A small group of authors, however, continued in the line of Weyl-geometric studies (Nishino/Rajpoot, Quiros, Ohanian et al.). Often they were not aware of the whole range of studies made in the 1970s to 1990s, and took up only one strand of the latter. With few exceptions,¹¹³ the majority of the aforementioned authors worked with two scalar

analysis indicate a way out of this impasse? – Warning: The mass-like expressions for W and Z in (11.87) cancel in $tr T_\Phi$ (Drechsler 1999, eqn. (3.55)) like in the energy-momentum tensor of the W and Z fields themselves.

¹¹²See, e.g., (Franklin 2017).

¹¹³For an exception, still standing under the spell of Drechsler/Tann though with a consistently scale-covariant approach and no explicit scale symmetry breaking term, see (Scholz 2011a).

fields, a *Higgs-like* one Φ with values in a spin $\frac{1}{2}$ representation of the electroweak group, and a real-valued one, here denoted by ϕ , the *gravitational scalar field*. In view of our main theme, we shall concentrate on the last group of authors, those who worked within the framework of Weyl geometry.

The modified Hilbert-Weyl term and kinetic terms of the scalar fields used by these authors were, up to notational conventions, of the form

$$L_{HW} = -\frac{\epsilon_{sig}}{2}(\zeta_1\phi^2 + \zeta_2\Phi^\dagger\Phi)R, \quad (11.88)$$

$$L_\phi = \epsilon_{sig}\frac{\alpha_1}{2}D_\nu\phi D^\nu\phi, \quad L_\Phi = \epsilon_{sig}\frac{\alpha_1}{2}(D_\nu\Phi)D^\nu\Phi^\dagger, \quad (11.89)$$

$$\mathfrak{L} = L\sqrt{|g|}, \quad \epsilon_{sig} = \begin{cases} +1 & \text{for } sig = (+ - - -) \\ -1 & \text{for } sig = (- + + +) \end{cases} \quad (11.90)$$

(in most cases $\alpha_1 = \alpha_2 = 1$), with Weyl-geometric scalar curvature R and electroweak and Weyl-geometric covariant derivatives D_μ .¹¹⁴ A quadratic curvature term L_{R^2} was added by some, but not by all authors. Yang-Mills terms of the electroweak connections (potentials) W for the SU_2 -component, B for hypercharge $U(1)$, and φ for the scale connection with field strength $f = d\varphi$, were also added,

$$L_{YM} = -\frac{1}{4}(\text{tr}(W_{\mu\nu}W^{\mu\nu}) + B_{\mu\nu}B^{\mu\nu} + f_{\mu\nu}f^{\mu\nu}). \quad (11.91)$$

Similarly Dirac kinetic terms $L_{\psi\text{kin}}$ and Yukawa mass terms $L_{\psi Y}$ for the different fermions ψ_{ih}^{fg} , with indices taking care for the various types and properties (fermion type $f = q, l$ for quark or lepton, generation $g = 1, 2, 3$, $i = u, d$ (“up, down”) for the 3-component of weak isospin, and helicity $h = R, L$) were adapted in a form compatible with the Weyl-geometric framework. The Dirac terms could be written with or without the Weyl-geometric scale connection term because, even if included, it finally cancels in the total expression. This had been noticed already by Hayashi and Kugo (see Section 11.3.2).¹¹⁵ We need not reproduce the explicit form of the fermionic terms here, but have to keep in mind that the Yukawa terms contained the matrices with relative mass coefficients (“mass matrix”) of the SM and a *scale-covariant* Higgs field.¹¹⁶

¹¹⁴In the high energy physical context, and accordingly in our section 3, signature of $g = (+ - - -)$. For sign conventions regarding curvature see footnote 31.

¹¹⁵See also (Blagojević 2002, p. 81).

¹¹⁶For an explicit form of Dirac kinetic terms and Yukawa mass terms see, e.g., (Nishino/Rajpoot 2009, eqn. (1.2)).

Breaking of scale invariance without an explicit mass term of the Higgs field became the crucial point for these authors. Because of its scaling behaviour ($w(\phi) = -1$), the gravitational scalar field ϕ already *specifies a preferred scale* in which it assumes a constant value ϕ_o (scalar-field gauge in the terminology of Section 11.2.1):

$$\phi(x) \doteq \phi_o = const \quad (11.92)$$

This was the reason why Utiyama called ϕ a “measuring field” already in the 1970s.

But the question still remains how such a specification – at first sight only a mathematical finding – can be incorporated into the material structures lying at the basis of measuring processes. In the context of the search for a connection between gravity and the *ew* sector of fundamental fields it seemed natural to search for a relation between the two scalar fields ϕ and Φ . For this a biquadratic/quartic potential in the two scalar fields, and a corresponding Lagrange term, plays a crucial role. Using the abbreviation $|\Phi|^2 = \Phi^\dagger \Phi$ this potential took the form:

$$V(\Phi, \phi) = \frac{\lambda_1}{4} |\Phi|^4 - \frac{\mu}{2} |\Phi|^2 \phi^2 + \frac{\lambda'}{4} \phi^4 \quad (11.93)$$

$$= \frac{\lambda_1}{4} \left(|\Phi|^2 - \frac{\mu}{\lambda_1} \phi^2 \right)^2 + \frac{\lambda}{4} \phi^4, \quad \lambda = \lambda' - \frac{\mu^2}{\lambda_1} > 0,$$

$$\mathcal{L}_V = -V(\Phi, \phi) \sqrt{|g|}. \quad (11.94)$$

Chromodynamics was usually not considered; this group of authors concentrated mainly on the electroweak sector of the SM and its possible link to gravity.

11.5.3.1 Nishino/Rajpoot

In 2004 two theoretical high energy physicists at California State University, *Hitoshi Nishino* and *Subhash Rajpoot*, set the goal of “extending the standard model with Weyl’s scale invariance”, adding that the scale invariance is “badly broken” at the order of the Planck mass/energy. They also made clear that according to the “philosophy advocated in the present work ... the standard model Higgs is not eliminated, and is the sought for particle” (Nishino/Rajpoot 2004, 1).

For adapting the fermionic fields to the differential geometric setting, the authors outlined the usual spinor calculus in a Weyl-geometric approach with scale dependent tetrads consisting of point-dependent bases $e_a = e_a^\mu \partial_\mu$ ($a = 0, \dots, 3$) and their dual forms, here denoted by $\vartheta^a = \vartheta_\mu^a dx^\mu$, and the metric $g_{\mu\nu} = \vartheta_\mu^a e_{a\nu}$. With $g(x) \mapsto \tilde{g} = e^{2\Lambda(x)} g(x)$ the tetrads have to be rescaled as follows:

$$\vartheta_\mu^a \mapsto \tilde{\vartheta}_\mu^a = e^{\Lambda(x)} \vartheta_\mu^a \quad e_a^\mu \mapsto \tilde{e}_a^\mu = e^{-\Lambda(x)} e_a^\mu, \quad (11.95)$$

that is $w(\vartheta^a) = 1$, $w(e_a) = -1$. The authors developed the Weyl-geometric affine connection, the corresponding spin connection, Weyl-geometric covariant derivatives, and curvature expressions.¹¹⁷

Utilizing this setup, they described a *two stage process* of symmetry breaking. In the first step they dealt with breaking the scale symmetry, formulated in terms of the compactified scaling group $\tilde{U}(1)$. The breaking was expressed “by setting” the value of the gravitational scalar field to a constant ϕ_o

$$\phi(x) = \phi_o \quad \text{with} \quad \zeta_1 \phi_o^2 = (8\pi G)^{-1}, \tag{11.96}$$

which in our terminology can be seen as the introduction of the scalar field (Einstein) gauge. In the second step, the *ew* symmetry was assumed to be broken “spontaneously”, as in the special-relativistic SM case ($SU_2 \times U(1)_Y \mapsto U(1)_{em}$). For the first step they gave a physical interpretation which bears some analogy with the Higgs “mechanism”:

At this stage the scalar field σ [here denoted ϕ , E.S.] becomes the Goldstone boson The vector particle associated with $\tilde{U}(1)$ breaking, the Weylon, absorbs the Goldstone field and becomes massive with mass M_S given by $M_S = \sqrt{\frac{3f^2}{4\pi G_N}} \approx 0.5 \times f M_P$ [f a coupling constant of the scale connection, E.S.]. (Nishino/Rajpoot 2004, 4)

Thus, the quartic potential (11.93) is reduced to the Higgs potential, much like in the SM with a cosmological term $\frac{\lambda'}{4} \phi_o^4$. In the ground state of the Higgs field only the cosmological term survives and the transition to scalar-field gauge endows the Higgs field with mass

$$m_H \doteq \sqrt{\mu} \phi_o. \tag{11.97}$$

After a short outline of how to adapt the parameters to the mass generation scheme of the SM, the authors concluded:

Our contention is that the present model presents a viable scheme in which gravity is unified, albeit in a semi-satisfactory way, with the other interactions. (. . .) When the complete theory of all interactions is found, the model in its present form, it is hoped, will serve as its low energy limit.

To conclude, we have accommodated Weyl’s scale invariance as a local symmetry in the standard electroweak model. This inevitably leads to the introduction of general relativity. (Nishino/Rajpoot 2004, 8)

This paper remained in a preprint stage; its content seems to have been presented at different conferences, but it never was published in a scientific journal. The reason

¹¹⁷The expression for the scalar curvature is given in the paper (and also in the later papers by the same authors) as $R = {}_gR - 6\nabla_\mu\varphi^\mu + 6\varphi_\mu\varphi^\mu$, where a coupling constant f introduced by the authors is here set to $f = 1$ and transcribed into our notation, (Nishino/Rajpoot 2004, eqn. (14)). The Weyl-geometric value (11.2) in our sign conventions would be $R = {}_gR - 6{}_g\nabla_\mu\varphi^\mu - 6\varphi_\mu\varphi^\mu$, cf. (Weyl 1918c, p. 21), (Drechsler/Tann 1999, eqn. (A 31)) and others. Nishino and Rajpoot apparently used inverted sign conventions for the curvature and the scale connection.

may have been that the authors considered it only as a first, provisional step. In the following years they extended their approach to a $SU(5)$ grand unified theory (GUT) (Nishino/Rajpoot 2007) and revised their presentation by taking up an idea going back to E. Stueckelberg (Nishino/Rajpoot 2009, 2011).

In the late 1930s *Ernst Stueckelberg* had been investigating the possibility of an interaction between a scalar field with nucleons. In this context, he introduced a massive scalar field B complementing an $U(1)$ potential A_μ , which expressed a field of electromagnetic type but with mass (i.e. similar to a Proca field). B was given a peculiar gauge behaviour involving a mass parameter m under $U(1)$ gauge transformations:

$$A_\mu \mapsto \tilde{A}_\mu = A_\mu + \partial_\mu \Lambda(x) \quad B(x) \mapsto \tilde{B} = B(x) + m \Lambda(x). \quad (11.98)$$

Transformations of type (11.98) were taken up by Pauli and others. They came to be known as *Stueckelberg transformations* and B was called the *Stueckelberg (compensating) field*. With an appropriate Λ , the Stueckelberg field allows one to specify a special gauge with $\tilde{B} = 0$ but *without breaking* the $U(1)$ symmetry, which is only given a “different realization” (in Drechsler’s terms quoted above, p. 317). This turned out to be crucial for the renormalizability of the theory and made the “Stueckelberg trick” attractive for quantizing the electromagnetic field or its relatives, the Proca-like fields.¹¹⁸

The careful reader may have noted the kinship between the Stueckelberg “trick” for $U(1)$ and the Higgs “mechanism” for the electroweak group. So did Nishino and Rajpoot. Moreover, they realized that, just by taking the logarithm, the transition to the Weylian scalar-field gauge can be given the form of a Stueckelberg transformation. Transliterated to our notation, they introduced an exponential expression of the form¹¹⁹

$$\phi(x) = \zeta_1^{-\frac{1}{2}} M_P e^{M_P^{-1} \beta(x)}. \quad (11.99)$$

Then the scale gauge transformation $\phi \mapsto \tilde{\phi} = e^{-\Lambda} \phi$ is expressed by

$$\beta \mapsto \tilde{\beta} + M_P \Lambda, \quad (11.100)$$

and the transition to the scalar-field gauge corresponds to $\tilde{\beta} = 0$, exactly like the Stueckelberg “trick”.

¹¹⁸The non-broken $U(1)$ symmetry is important for the BRST relations, the quantum analogue of the Noether relations. See (Ruegg et al. 2003, 75ff).

¹¹⁹The factor $\zeta_1^{-\frac{1}{2}}$ in (Nishino/Rajpoot 2009, eqn. (2.1)) was set by them to $\zeta_1 = 1$ while transforming the Lagrangian into their eqn. (2.3). The follow up paper (Nishino/Rajpoot 2011, second paragraph of section 2) shows that this reduction was intended. Of course, a different factor ζ_1 would heavily influence the mass calculation in (11.101).

Nishino and Rajpoot thus rewrote their basic Lagrange density equivalent to our equations (11.88, 11.89, 11.91, 11.93) in terms of the logarithmized scalar field (Nishino/Rajpoot 2009, eqn. (2.3)) and normed it to scalar-field gauge (Nishino/Rajpoot 2009, eqn. (2.6). Then they could read off the mass expression m_φ for the scale connection field (“Weyl field”).¹²⁰ In scalar-field gauge the kinetic terms (11.89) of ϕ and Φ acquire forms which make them contribute to m_φ . For ϕ this is

$$\frac{1}{2}D_\nu\phi D^\nu\phi \doteq \frac{1}{2}(fM_p)^2\varphi_\nu\varphi^\nu, \quad (11.101)$$

while for Φ the contribution to the mass of φ is $f(\Phi^\dagger\Phi)$ (after *ew symmetry breaking* fv^2 , with v^2 the vacuum expectation value of the operator $\Phi^\dagger\Phi$). In any case the contribution due to Φ is much less than the one from ϕ and from the modified Hilbert term (11.88), both of which are on the order of the Planck scale. Thus it may be safely neglected at several orders of magnitude.¹²¹

In the imaginative language of the elementary particle community, Nishino and Rajpoot commented that the scalar field is “now eaten up by the Weylon”. A little later they added, more technically: “After all, the Weylon \hat{S}_μ [our φ_μ , E.S.] acquires the mass fM_p , the compensator φ [our β , E.S.] is absorbed into the longitudinal component of \hat{S}_μ , and the potential terms are reduced to the Higgs potential in SM ...” (Nishino/Rajpoot 2009, 3). With this explanation they clad the mass derivation for the scale connection field in the mantle of a narrative which is widespread in their community and usually accepted as scientifically founded.¹²²

In a follow up paper, the two California State physicists came back to this topic, for which they presented some results concerning a quantized version of their theory. They started from a Lagrangian given in terms of the logarithmized scalar field (Nishino/Rajpoot 2009, eqn. (2.3)) and with modified Hilbert term

$$L_{HW} = -\frac{1}{2}\left(\zeta_1 M_P^2 e^{2M_P^{-1}\beta(x)} + \zeta_2 \Phi^\dagger\Phi\right) R, \quad (11.102)$$

(R being the Weylian scalar curvature, written there as \tilde{R}). At this point Nishino and Rajpoot left the track of Weyl geometry and decided to switch to the JBD paradigm. They considered the initial Lagrangian a “Jordan frame” and wanted to transform it to an “Einstein frame”.¹²³ For this goal they performed a “Weyl

¹²⁰Warning: Nishino/Rajpoot used the notation φ for the Stueckelberg “compensator”, i.e. our β , and S_μ for the scale connection (the potential of the “Weyl field”), our φ_μ . In order to avoid confusion the notation in the present paper has been homogenized for the authors discussed here.

¹²¹Nishino/Rajpoot did not consider the contribution of the modified Hilbert term, in contrast to Smolin and Cheng (see Section 11.5.1).

¹²²Compare (Stoeltzner 2014).

¹²³Strictly speaking, their framework does not contain any meaningful “Jordan frame” because their Weyl structure is not integrable, and thus the purely Riemannian representation of the affine

rescaling for the vierbein or metric only”, i.e. a “*field re-definition*” which did not include the corresponding transformations of the scale-covariant fields and the scale connection.¹²⁴ Referring to calculations in the framework of JBD theory, they arrived at a reduction of the Hilbert term to a form which depends only on the Riemannian component ${}_g R$ of the scalar curvature. According to their calculation, the scale connection contributions drop out of the Lagrangian (but not the Yang-Mills term for the scale curvature). In other words, they achieved a reduction to the Einstein frame form of JBD with two scalar fields and an additional Yang-Mills field (Nishino/Rajpoot 2011, eqn. (2.10)).

On this basis, the authors performed a series of calculations at the quantum level. They derived various anomalies (Adler-Bell-Jackiw and trace), studied the possibility for cancelling the remaining (trace-) anomalies, considered quantum corrections to the cosmological constant, and studied the perturbative renormalizability of their model and possible new divergences. All in all, these were remarkable results; but they were arrived at using a hybrid approach which started in a setting of Weyl-geometric gravity and ended in JBD gravity, after performing an artificial and methodologically unconvincing transition by a “field re-definition” type of rescaling. In spite of such shortcomings, these derivations were a notable step towards connecting the electroweak sector of elementary particle physics with gravitational structures, mainly formulated in a Weyl-geometric framework.

11.5.3.2 Hao Wei, Rong-Gen Cai, Quiros

H. Nishino and S. Rajpoot were not the only researchers who thought about how to establish a connection between gravity and the SM fields by exploiting Weyl-geometric methods. Even though we have to be selective here, it should be clear that the Weyl-geometric approach continues to be relevant in the era of the Higgs boson (or some close relative) found in experimental observations. A talk given in July 2004 by Hung Cheng at the Institute for Theoretical Physics of the Chinese Academy of Science in Beijing seems to have initiated interest in Weyl-geometric methods by Chinese theoretical physicists *Hao Wei, Rong-Gen Cai* and others.¹²⁵ It was natural for them to take the “Cheng-Weyl vector field” (i.e., the Weylian scale connection with massive boson studied by Cheng in the late 1980s) and Cheng’s view as their starting point for a new look at the standard model of elementary particle physics (Wu 2004; Cai/Wei 2007).

connection presupposed for an ordinary Jordan frame does not exist. The Einstein frame, on the other hand, is meaningful in any Weyl-geometric gravity approach with a scale-covariant scalar field and corresponds to scalar-field gauge (11.92).

¹²⁴“Note that the Weyl rescaling we made is a field re-definition, but it is not a part of any local scale transformation which is defined to act not only on $g_{\mu\nu}$ but also on Φ and φ as in (2.2) [the equation for the full gauge transformation, E.S.]” (Nishino/Rajpoot 2011, p. 4).

¹²⁵(Cai/Wei 2007, Acknowledgments)

Another road was taken by *Israel Quiros*, who was located at the time of our interest in Guanajuato, Mexico. Coming from a background in Jordan-Brans-Dicke gravity and cosmology (see Section 11.6.2.2), he developed thoughts of his own about how “scale invariance and broken electroweak symmetry may coexist together” (Quiros 2013). In this conceptually clear paper he gave a concise introduction to the basic ideas of integrable Weyl geometry. Furthermore, he showed that by using Weyl-geometric gravity not only can the scale covariance of the SM fields be imported into a general relativistic framework but it can even be upheld after breaking the ew symmetry. One only need to accept the use of mass parameters m which scale with weight $w(m) = -1$.

For his presentation Quiros used a simplified version of the Lagrangian (11.88ff.) similar to the one of Nishino/Rajpoot, whose papers he probably did not yet know. He encoded the gravitational scalar field in terms of a point-dependent scalar exponent of the factor in the Hilbert-Weyl term, written by him as φ (in order to avoid confusion, we shall transliterate it like above as β). Using our notation, he wrote

$$\zeta_1 \phi(x)^2 = M_p e^{\beta(x)} \quad (11.103)$$

and considered Weylian scale connections exclusively of the form

$$\varphi = \varphi_\mu dx^\mu = d\beta \iff \varphi_\mu = \partial_\mu \beta \quad (11.104)$$

(Quiros 2013, eqn. (8)). This implies the restriction

$$\phi = const \iff \beta = 0 \Rightarrow \varphi = 0. \quad (11.105)$$

In our terminology (11.103) implies an *identification of the Riemann gauge and Einstein gauge*. This probably went unnoticed by the author. For the basically geometrical and conceptual purposes of this paper such a restriction had perhaps no great disadvantage, although the dynamical role of the scalar field was trivialized by this specialization.

11.5.4 Towards Weyl Scaling at the Quantum Level

11.5.4.1 Scale-Invariant Quantization Procedures

Problems of a more fundamental nature have been posed by a group of theoretical physicists working in Trieste. *Alessandro Codello*, *Giulio D’Orico*, *Carlo Pagani* and *Roberto Percacci* recently reconsidered the question of how scale invariance behaves under quantization if one approaches this by using the so-called “renormalization group” (RG) and functional integral methods. In the report on their work (Codello et al. 2013) they rebutted the general view that quantization necessarily

leads to a breaking of (point-dependent) scale symmetry even if the classical Lagrangian is scale invariant. In a step by step argument they showed how the functional integrals can be given a scale-invariant form by using an integrable Weyl-geometric background and a gravitational scalar field χ of weight $w(\chi) = -1$, called a “dilaton”, while leaving the external fields unquantized at the first stage.

The authors’ basic idea was that “one can make any action Weyl-invariant by replacing all dimensionful couplings by dimensionless couplings multiplied by the powers of the dilaton” (Codello et al. 2013, p. 2). Then a dimensional coupling coefficient of scaling dimension k , let us say μ , is turned into a coupling parameter of the form $\chi^{-k}\check{\mu}$ with a “dimensionless”, i.e. non-scaling, constant $\check{\mu}$. The authors achieve scale covariance/invariance of the fields, respectively actions, by using a Weyl-geometric expression with respect to an integrable scale connection with coefficients

$$b_\mu = -\chi^{-1}\partial_\mu\chi \quad (\text{Codello et al. 2013, p. 3})^{126}. \quad (11.106)$$

Like Nishino and Rajpoot, they considered this as a kind of gravitational equivalent of the “Stückelberg trick”. Their main work then consisted in showing that Weyl invariance, which is easily achievable for the classical action, is left intact within their framework for the functional integrals, the differential equation governing the renormalization flow equation, and the UV and IR endpoints of the flow.

The energy-momentum tensor of (pseudo-) classical quantum matter fields (scalar or Dirac spinors) will have a vanishing trace, whereas the expectation value of the quantized trace no longer vanishes. This so-called *trace anomaly* of quantization has puzzled theoretical physicists for a long time, and has usually been interpreted as a breaking of scale invariance at the quantum level. The Italian authors came to a different conclusion. They explained that, although the “trace anomaly” is still present in their approach, it *no longer signifies breaking of the local scale invariance*. The reason lies in a cancellation of the trace terms of the quantized fields by corresponding counter-terms that arise from the scalar field, the “dilaton” in the language of the paper.

After some comments on the quantization of the metric field, and further discussions of the difference between strictly conformal theories and the Weyl-geometrically “conformalized” ones, the authors ended with the remark: “The present work provides a general proof that with a suitable quantization procedure, the equivalence between conformal frames can also be maintained in the quantum theory” (Codello et al. 2013, p. 21). But they also stated clearly that their quantization procedure does not lead to new physical effects. In this sense their research shows a certain analogy with Kretschmann’s view of diffeomorphism invariant reformulations of physical theories which do not per se lead to new physical insights. Even so, the authors succeeded in showing that the extension of the mathematical automorphism group of the underlying theories (SM fields, implicitly also gravity

¹²⁶Note that the differential form $b = b_\mu dx^\mu$ is sign inverted in comparison with our conventions of Section 11.2.1.

theory) can be upheld under quantization. Whether a further enrichment of the theories delivers new insights at the quantum level will be a question for the future. Probably this can only be the case if the scalar field and/or the scale connection acquires a dynamical role beyond its purely mathematical “compensatory” character in the scale transformation.

11.5.4.2 Ohanian’s Retake of a “Spontaneous” Breaking of Symmetry

Such an attempt at giving the scale connection a dynamical role has been made by *Hans Ohanian* from the University of Vermont. He proposed a model which connects the standard model fields with general relativity in a Weyl-geometric framework, in which a complex scalar field χ (“dilaton”) acts as the crucial mediator. This undergoes spontaneous breaking of local scaling symmetry, which the author preferred to call conformal symmetry,¹²⁷ by a mechanism very similar to the breaking of electrodynamic $U(1)$ symmetry in a model studied by (Coleman/Weinberg 1973). If gravitational effects can be neglected, Ohanian’s adaptation leads to the SM field content in flat spacetime (Ohanian 2016). If, on the other hand, gravity is taken into account, the transition from quantum to classical matter being leapfrogged, it leads to Einstein gravity as an “effective field theory”. Regarding the conformal expression of fields Ohanian used a “conformalization” procedure with additional terms in the (Riemannian) scalar curvature (in place of the more natural Weyl-geometric expressions). Ohanian proposed to assimilate the result of Coleman/Weinberg by a simple substitution of coefficients. He then concluded: “After symmetry breaking, neither the scalar field nor the vector field reveal themselves at the macroscopic level, and we can ignore the effects of the Weyl gauge-vector on the transport of lengths . . .” (Ohanian 2016, 10f). Because of the conformal coupling of the scalar fields to the *Riemannian* scalar curvature, Ohanian found that in his approach a *modification of Riemannian geometry is excluded* in the long-range regime, about which he commented: “This is in contrast to the standard Brans-Dicke theory, in which the massless scalar field makes a contribution to long-range gravitational effects . . .” (ibid.).

In the high energy, short-range regime Weyl-geometric curvature does play a role in this model, as Ohanian discussed in section 4 of his paper. Then the scale connection constitutes a “vector” field of its own, similar to the electromagnetic field but with a mass term and with the dynamical current of the scale symmetry $\mathfrak{J}^\mu = \frac{\partial \mathcal{L}}{\partial \varphi_\mu}$ as right hand side of the dynamical equation.¹²⁸

Ohanian conjectured that certain problematic features in the purely conformal approaches are essentially due to the lack of a coherent metrical structure. In Weyl

¹²⁷Ohanian preserved the label “scale transformation” for a global usage in Minkowski space, where, in addition to the rescaling of the fields $X \mapsto \tilde{X} = \Omega^k X$, a space dilation $x \mapsto \tilde{x} = \Omega x$ is applied (Ohanian 2016, p. 25).

¹²⁸ $\partial_\nu (\sqrt{|g|} f^{\mu\nu}) = \mathfrak{J}^\mu$. In Ohanian’s Lagrangian φ couples only to the “dilaton” scalar field χ . This leads to a form for the variation of the Lagrangian under scale transformations such that the dynamical current coincides with the Noether current (Ohanian 2016, eqn. (13)).

geometry the scale connection serves to make a Weylian metric consistent with conformal rescaling. Ohanian therefore ended his paper with a remark that goes right to the heart of this matter: “If the analysis of Ehlers et al. is correct, the absence of a Weyl vector and its geometric paraphernalia is a fatal mistake – if no Weyl vector, then no conformally invariant theory with a geometric interpretation” (Ohanian 2016, p. 16). Ohanian’s proposed model indicated why and how a Weyl field with curvature at the short-range, high energy level loses its curvature in the low energy regime, leading to Einstein gravity in the long-range limit.

Ohanian, like many other authors, perceived the transition between the energy regimes (high – low) mainly, though not exclusively, as *successive temporal stages* in the hypothesized cosmic development. This view fits well with the mainstream narrative connecting cosmology and high energy physics shortly after the big bang. A philosophically inclined reader may notice that one also can interpret this kind of transition non-temporally, as a *structural passage between different energy levels*, present at any time and any place of the world. Such a passage may be of importance, independent of the view regarding the reality content of the big bang picture. Physicists may well claim that, e.g., the LHC experiments are important because they explore what the world looked like a few “nanoseconds after the big bang”. But one need not take such stories at face value in order to appreciate the activities aimed at gaining knowledge about the respective energy levels and the transitions between them.

11.6 Weyl-Geometric Models in Astrophysics and Cosmology Since the 1990s

11.6.1 *The Broader Context: Scalar Fields in Gravity, Conformal Rescaling*

In the 1970s JBD theory underwent a conflicted development. On the one hand, the increasing precision of radar tracking observations within our planetary system confirmed that Einstein’s theory alone gives an extremely accurate description of gravitational phenomena.¹²⁹ A tentative modification of the latter by a Brans-Dicke type scalar field was therefore seen as superfluous, at least on this level. For conventional astronomical purposes this would either involve an extremely high value of the coupling coefficient ξ of the kinetic term in (11.10) if JBD were to be adequate at all. On the other hand, the rise of particle cosmology as a new subfield of theoretical physics opened ample room for studying models in an assumed very early phase of the universe. Here it appeared reasonable to think about modified gravity and elementary particle physics as an ensemble. A fertile

¹²⁹See C. Will’s contribution to this volume.

environment for studying speculative models thus emerged, some of which were designed for combining the gravitational scalar field and a Higgs-type scalar field from elementary particle physics (Kaiser 2006, 2007). This gave new motivations and incentives for studying scalar-tensor theories completely different from those of the 1960s and 70s (Capozziello/Faraoni 2011, chaps. 3, 7), (Clifton/Ferreira et al. 2012).

One of the new roles rehearsed for the scalar field on this stage was that of an agent, called *inflaton*, which drives a hypothetical phase in a very early accelerated expansion of spacetime. Another role arose from string theory, where a new type of scalar field, a so-called *dilaton*, entered the stage. Originally it coupled to the trace of the (2-dimensional) stress tensor of the string; but the dilaton re-appeared in the form of a constraint for restoring conformal symmetry. After its breaking under quantization, it acted as a source term in a classical Einstein-like equation. This gave rise to speculations aimed at deriving Einstein gravity as an effective theory arising from string theory with the dilaton scalar field and conformal symmetry acting as mediators (Brans 2005, p. 14f).

All in all, a vast field for studying scalar field theories in connection with generalized theories of gravity arose.¹³⁰ Only a few authors who took up work in this field employed Weyl geometry for their purposes. This was the case, e.g., in string models, but that line of research lies outside the scope of this survey and would require a separate study of its own. Here we focus only on the more mundane manifestations of Weyl geometry in cosmology and astrophysics during the last two decades. Because of the close kinship between Weyl-geometric rescaling and conformal invariance of field theories in a Riemannian environment, I begin with a few examples of recent conformal approaches in cosmology. These are, of course, far from exhaustive, but they were selected because they connect in specific ways to our core topic.

11.6.1.1 Conformal Approaches in Cosmology

An unusual analysis of the “dark” sectors that have surfaced in recent cosmology was given by *Philip Mannheim* and *Demosthenes Kazanas*. They argued that the flat rotation curves of galaxies can be explained on the basis of a conformal approach to gravity (Mannheim 1989). In their theory, a static spherically symmetric matter distribution was described by the solution of a fourth-order Poisson equation

$$\nabla^4 B(r) = f(r), \quad (11.107)$$

where a typical coefficient $B(r)$ is proportional to $-g_{oo} = g_{rr}^{-1}$ in a metric $ds^2 = g_{oo}dt^2 - g_{rr}dr^2 - r^2d\Omega^2$ (up to a conformal factor). The r.h.s. of the Poisson equation $f(r)$ depends on the mass distribution, e.g. in a spiral galaxy.

¹³⁰For extensive surveys of this field see (Fujii/Maeda 2003; Capozziello/Faraoni 2011).

A comparison of results from their theory with data from eleven galaxies whose rotation curves exhibit different behaviours led to a good fit. This encouraged the authors to present their approach as a candidate for a modified gravity that could explain dark matter phenomena. (Mannheim 1989, 1994, 2012).

During the following years, Mannheim extended this approach to the question of dark energy. In the special case of conformally flat models, such as Robertson-Walker geometries, he proposed taking the Hilbert-Einstein Lagrangian term $-\frac{1}{12}|\phi|^2 R\sqrt{|det g|}$ of a conformally coupled scalar field ϕ as part of the *matter* Lagrangian. Due to this sign choice, he arrived at a version of the Einstein equation with *inverted* sign. He interpreted this as a kind of “repulsive gravity”, which supposedly operates on cosmic scales in addition to the “attractive gravity” on smaller scales, as indicated by the conformally modified Schwarzschild solution. This repulsive form of gravity might then take the place of the dark energy corresponding to the cosmological constant term in standard gravity (Mannheim 2000, 729). In spite of such a drastic alteration of Einstein gravity, Mannheim did not consider this conformal viewpoint to disagree with the standard model of cosmology and its accelerated expansion. Rather he argued that his approach could lead to a more satisfying explanation for the expansion dynamics with “repulsive gravity” taking over the role of dark energy. Moreover, he expected that a conformal approach with quadratic curvature terms ought to shed new light on the initial singularity and, perhaps, also on the black hole singularities inside galaxies.

A completely different approach using local conformal symmetry in particle physics and cosmology is due to *Izhak Bars, Paul Steinhardt and Neil Turok*. A silent background for their interest in this question seems to have been the idea of a cyclic, respectively oscillating, model of the universe, which was proposed a decade earlier by two of them as an alternative to the “inflationary” paradigm (Steinhardt/Turok 2002). In this alternative proposal, the minima of the oscillation were related to some kind of speculative physics of the string and brane type.¹³¹ In (Bars et al. 2014) they explored the possibility that a conformal theory of gravity and the standard model fields might suffice for understanding the bridging process between two cycles without necessarily invoking new speculative assumptions. They worked with a locally scale invariant version of the standard model combined with gravity, similar to Nishino/Rajpoot (Section 11.5.3) but in the framework of purely conformal geometry rather than Weyl geometry. They considered a complex-valued gravitational scalar field ϕ , called a *dilaton*, in addition to the Higgs field Φ , both scaling with the same weight (in our notation $w = -1$). The dilaton couples only to the Higgs field by a common biquadratic potential as in (11.93) and to the right-handed singlet neutrinos by Yukawa terms of its own (Bars et al. 2014, p. 6). All other masses are “generated” by coupling to the Higgs field following the standard model.

¹³¹For a historical discussion of oscillating models see (Kragh 2009) and, in an even wider perspective, H. Kragh’s contribution to this book.

The authors investigated possible general forms for locally scale-invariant gravitational Lagrangians, including a kinetic term for the dilaton (eqn. (10), loc. cit.). They were for several reasons heading towards “a fully scale-invariant approach to all physics” (p. 5, loc. cit.). At first, the “dimensionless constants in a conformally invariant theory are logarithmically divergent as opposed to the quadratic divergence of a bare Higgs mass term”, but the recent studies of (Codello et al. 2013) have shown that “the local scale invariance survives even though there is a trace anomaly” (Bars et al. 2014, p. 2). Moreover, so they claimed, the conformal freedom for choosing different scale gauges made their cosmological models geodesically complete. That was a bit cavalier, but it is not the aim of this paper to evaluate such claims critically.¹³² More important here is to recognize the similarity in outlook between this work and the Weyl-geometric proposals for combining gravity with standard model fields in a strictly scale-invariant approach. Here the possibility of a putative “geodesic completeness” of cosmological models came into play, although in a rather peculiar way that did not take into account the problem of an invariant characterization of the proper time along timelike geodesics.

This question was discussed in more detail by *R. Penrose* in his recent proposal for embedding the standard model of cosmology in a long cycle of iterations connected by conformal bridges between Riemannian phases of cosmic evolution (Penrose 2006). He argues that for very high energy states in the past timelike trajectories lose their physical meaning, so the only physically relevant information is carried by the structure of the lightcones. By some not yet understood processes, a similar argument is imputed for states in the asymptotic future. This idea develops a purely conformal perspective indicating how Riemann-Einstein gravity can be extended beyond the conformally compactified past and future infinities, which would probably fit well with the Bars/Steinhardt/Turok approach. Both proposals assume that it is possible to develop a meaningful physics for the bridging process between two cycles using an approach that neglects all those geometrical features which distinguish Weyl geometry from a purely conformal structure.

11.6.2 Diverse Views of Weyl Geometry in Cosmology

11.6.2.1 Continuation of Rosen’s Work

M. Israelit investigated Weyl-geometric methods in cosmology during the first half of the 1990s together with his mentor N. Rosen. After Rosen’s death in 1995 he continued publishing on his own for nearly two more decades.¹³³ In this work

¹³²The authors declared geodesic incompleteness as “an artifact of an unsuitable frame choice: geodesically incomplete solutions in Einstein frame may be completed in other frames, even though the theories are entirely equivalent away from the singularity” (Bars et al. 2014, p. 13).

¹³³Israelit died in 2015 at the age of 87. His last paper known to me is (Israelit 2012), a slightly changed version of (Israelit 2010).

he studied the problem of dark matter from different perspectives, but always based on geometrical fields. In (Israelit/Rosen 1992) both authors explored the neutral massive boson interpretation of the Weylian scale connection, already hinted at by Rosen in his 1982 paper (cf. Section 11.3.2.2). The authors assumed a “chaotic Weylian microstructure”, constituted physically by a “Weylon gas”. At large distances the scale curvature effects were negligible so that a Riemannian space structure arose in their approach. On this basis Rosen and Israelit studied a hypothetical Bose-Einstein *Weylon gas* satisfying the equation of state $\rho = 3p$, while drawing consequences for different cosmological models in Einstein gravity (Israelit/Rosen 1993).

In one of their next papers, they turned to the *scalar field* as a potential contributor to the dark matter “pervading all of cosmic space”. By this they meant on the largest possible scales, not in the sense of local inhomogeneities in galaxies (treated in theories of the MOND family) or in galaxy clusters. Although in their approach the Einstein gauge ($\beta = 1$) leads to “the usual formalism of general relativity” (Israelit/Rosen 1995, p. 764), the authors believed that different gauges with non-constant β -field might lead to new physical insight. They declared that “although the gauge function is arbitrary, it leads to the presence of dark matter which, in principle, can be observed” (Israelit/Rosen 1995, pp. 777). This was not particularly convincing and it remained an open question how such an observation “in principle” might actually be made.

In some papers from the late 1990s and several at the beginning of the new millennium Israelit continued this research line using *integrable Weyl-geometric gravity*. In this context he realized that even under the assumption of an integrable Weylian scale connection the resulting modification of Einstein gravity can be *nontrivial* if the potential of the scale connection w is different from the scalar field, respectively its logarithm (Israelit 1999b, chap. 7), (Israelit 1999a, eqs. (17)f.) (compare our eqs. (11.7), (11.9)). Israelit derived the dynamical equations with respect to $g_{\mu\nu}$, φ_μ and β , as well as the Noether relations due to diffeomorphism invariance and scale invariance of the Lagrange density. The latter relations showed that on shell of the Einstein equation the dynamical equations of the scale connection φ_μ and of the scalar field β are equivalent.¹³⁴

Israelit’s aim was not only to explain dark matter but also to account for the accelerated expansion of standard cosmology by the gravitational scalar field which he called the “Dirac gauge function”.¹³⁵ In these papers he tried to convince his colleagues that “cosmic matter was created by geometry”, viz. out of the energy of the gravitational scalar field (Israelit 2002a, p. 295). According to him, his scalar field was empowered to generate dark matter and the magical substrate of *quintessence* flourishing in the mainstream narratives on the early “history” of the

¹³⁴Because of scale invariance there is, in fact, only one true scalar field degree of freedom (compare Section 11.2.1).

¹³⁵(Israelit 1996, 1999a, 2002a,b); chapters 6 and 7 in his book (Israelit 1999b).

universe. These were imaginative proposals, but one may reasonably doubt that they stood on a solid base.

In one of his last papers, Israelit went back to considering a non-integrable Weylian scale connection, now no longer as a representative of the electromagnetic potential but rather as a field with massive bosons, “Weylons”, of spin -1 and mass $> 10 MeV$. On a microlevel, so he argued, the Weyl-geometric structure appears non-integrable, “chaotic”, while on larger scale there remains an effective gauge “vector field” with vanishing curvature. The author concluded by remarking that “the purpose of the present work was to show that on the basis of the Weyl-Dirac theory one can build up a model, where conventionally matter, DM and DE are created by geometry. This aim is achieved” (Israelit 2010, sec. 8).

The “creation” described by Israelit did not even claim to establish a connection between geometry and the standard model fields. His discussion appeared as a reflex far from the cosmological mainstream, which was dominated very successfully by elementary particle physicists with their debates over the “early history” of the universe (Kaiser 2006). Perhaps this is one of the reasons why we find so little overlap between the work of Rosen and Israelit and the investigations in the last section or the ones to be discussed in the next.

11.6.2.2 Weyl-Geometric Extensions of Gravity: Trivial or Provocative?

Coming from Jordan-Brans-Dicke theory, *Israel Quiros* became interested in Weyl geometry while still working at Santa Clara, Cuba, several years before the work described in Section 11.5.3.2. He was one of those in the JBD community who took Dicke’s proposal seriously, so he aimed to formulate natural laws so that they do not depend on (localized) choices of measurement units (Quiros et al. 2000; Quiros 2000a).¹³⁶ At first, he developed the formulation of “dual” views for the interchange from the Jordan to the Einstein frame (Quiros 2000b). A decade later, after he had moved to León, Mexico, he wrote a joint paper with three other authors, *José E. Madriz Aguilar, Ricardo García-Salcedo, Tonatiuh Matos*, in which they explained how the different frames of JBD theory may be interpreted as “complementary geometrical descriptions of [the] same phenomenon” (Quiros et al. 2013).¹³⁷

From there he was only a small step away from entering Weyl-geometric gravity. As seen in the last section, Quiros was looking for a common perspective on gravity and a scale-invariant formulation of SM model fields (Quiros 2014b). In a recent paper he had investigated the purely conformal approach to scale-invariant Lagrangian field theories, but concluded by criticizing its lack of a well defined metrical structure with a uniquely determined affine connection:

... that there will be problems with a theory which pretends to be Weyl-invariant only because the action – and the derived field equations – is invariant under (2) [point-dependent

¹³⁶Compare on this point (Capozziello/Faraoni 2011, pp. 86ff).

¹³⁷J.E. Madriz Aguilar was a student of C. Romero of the Brazilian school, see Section 11.6.4.

scale transformations, E.S.], but which is sustained by spacetimes whose geometrical structure does not share the gauge symmetry of the action. (Quiros 2014a, p. 3)

Quiros therefore pleaded for using Weyl geometry as an appropriate overall framework for his research goal.

But, alas, simplifying the Lagrangian used in (Quiros 2014b) he only introduced a kinetic term for the Higgs (or a Higgs-like) field Φ , not for the gravitational scalar field ϕ coupling to the Hilbert term. With a gravitational Lagrangian including a quartic potential

$$L_{grav} = \frac{1}{12}\phi^2 R + \lambda\phi^4 \quad (\text{Quiros 2014a, eqn. (20)}) \quad (11.108)$$

he found that his scalar field equation for ϕ reduced to the trace of the Einstein equation, just as in the case of conformal coupling in Riemannian geometry. After pondering the possibility of having “an infinity of feasible — fully equivalent — geometrical descriptions” and the resulting paradoxical picture of an “infinity of possible patterns of cosmological evolution” he passed over to the Einstein scalar-field gauge as the “simplest gauge one may choose”.

Choosing (11.108) as the gravitational Lagrangian resulted in a Hilbert action of Einstein gravity “minimally coupled to the standard model of particles with no new physics beyond the standard model at low energies” (Quiros 2014a, p. 9). This led to the simple observation that conformal rescaling allows one to scale singularities away, but without any new physical insights or effects; in this sense the Weyl-geometric extension of gravity considered by Quiros up to 2014 remained physically trivial. Still, he gave a conceptually clear exposition of ideas and methods, so we may hope that in the further development of Quiros’ research program he will go beyond these limitations.

Carlos Castro, after the turn of the millennium working at the Centre for Theoretical Studies of Physical Systems in Atlanta, USA, had become acquainted with Weyl geometry already in the early 1990s (see Section 11.4.2.1). At that time Santamato’s proposal for using Weyl’s scale connection for geometrizing the quantum potential stood at the center of his interests (Castro 1992). When he became aware of the new attempts to use Weyl-geometric methods in gravity and in high energy physics, he took up the Weylian thread again to pursue a new question: how might one use Weyl’s scale geometry to gain a deeper understanding of dark energy and also, perhaps, to explain the Pioneer anomaly, which was still a challenge for gravity theories at this time (Castro 2007, 2009).¹³⁸ In contrast with Quiros’ more sober perspective, Castro has tended to speculate about grand visions opened by his new-found interest in Weyl-geometric methods. An even sharper contrast arises

¹³⁸A few years later high precision numerical modelling showed that thermal effects can completely account for the observations known as the flyby anomaly of the Pioneer spacecrafts (Rievers/Lämmerzahl 2011).

from his expository style, which lacks a clear conceptual basis, and for this reason we need not discuss this work further here.

Another unconventional view was put forward by the present author (*Erhard Scholz*, Wuppertal). His historical studies on the work of Hermann Weyl led him to the idea that a comparatively simple modification of Riemannian geometry by integrable Weyl geometry, combined with a non-trivial scalar field extension of Einstein gravity (in the sense of our Section 11.2.1 with $v + w \neq 0$), might shed new light on certain points of present-day cosmology. He was pleased to discover some recent interest in Weyl-geometric gravity among members of the Munich “group”, although their work went in a different direction (Section 11.5.2).

Scholz also found it most intriguing to see that in a Weyl-geometric approach to gravity the cosmological redshift need no longer be regarded as due to an expansion of the spacelike folia of Friedmann-Robertson-Walker manifolds. This was clear because in the transition from the Riemann gauge to Einstein gauge the warp function may be scaled away partially or completely (Scholz 2005b,a).¹³⁹ Thus, part of the cosmological redshift z may be due to the time component φ_o of the scale connection rather than to a spatial expansion of the “universe”. The reason for this comes from the scale invariance of z , if scale-covariant geodesics of weight $w = -1$ are used. For cosmological observers following a timelike geodesic flow X the redshift is given by the quotient of energies E_o, E_1 of light signals (idealized “photons”) at the event of emission p_o and of observation p_1 :

$$z + 1 = \frac{E_o}{E_1} = \frac{g(\gamma'(\tau_o), X(p_o))}{g(\gamma'(\tau_1), X(p_1))}, \quad (11.109)$$

where $\gamma(\tau)$ denotes the null-geodesic representing the trajectory of the signal. Because of the parametrization of the geodesics with weight $w = -1$ the quotient is scale invariant. Although for Robertson-Walker models with non-trivial (i.e., non-constant) scalar field and warp function $a(t)$ in the Riemann gauge the *redshift* seems to result from an expanding warp function, in the Einstein scalar-field gauge this *may* at least partially be due to the scale connection, i.e., to a *field effect* of the additional component of the gravitational structure.

This effect is particularly striking in certain models which appear as expanding in the Riemann gauge, but have a static metric in the Einstein gauge (Scholz 2005a, 2009). In these the cosmological redshift in the Einstein gauge turned out to be completely due to the time component of the Weylian scale connection, $H = \varphi_o$, with H the Hubble parameter. Although Scholz initially overestimated the physical import of this example, it could probably be a fruitful epistemic provocation.¹⁴⁰ As a toy model it may thus continue to serve as an incentive for critically rethinking the foundations of our standard picture of the universe.

¹³⁹A similar argument was already given by Rosen (see Section 11.3.2.2) and more recently by (Romero et al. 2012, sec. 7); compare (Perlick 1989, chap. 5).

¹⁴⁰It was the topic of the author’s talk at the Mainz conference.

11.6.3 Attempts at Dark Matter, MOND-Like

Since the 1980s, the success of *modified Newtonian dynamics* (MOND) for explaining the rotation curves of galaxies and the Tully-Fisher relation between the luminosity of spiral galaxies and their angular velocity has led to diverse attempts to incorporate general relativistic features (Sanders 2010). Some of these introduced non-geometrical structures, like an additional vector field in the so-called tensor-vector-scalar field theory, TeVeS; but the earliest attempt at a relativistic MOND-like theory was formulated by *Mordechai Milgrom* and *Jacob Bekenstein* in the framework of JBD gravity (Bekenstein/Milgrom 1984).

This approach worked with a non-quadratic kinetic Lagrangian for the scalar field with MOND-typical transition function, hence its name “relativistic a-quadratic”: rAQUAL theory. Bekenstein and Milgrom were torn between the Jordan frame and Einstein frame. In a later review paper by the former, the Jordan frame was declared to be the “physical metric”, while the Einstein frame was considered as the “primitive metric” (Bekenstein 2004, p. 6).¹⁴¹ In this framework the MOND-like free fall of particles in an extremely weak gravitational field could be derived. This approach was not free of shortcomings, however, as the authors themselves remarked. For example, lensing effects seemed to be unexplainable because the conformal change between the two “dual” frames seemed not to affect light-like geodesics.¹⁴² Moreover, the scalar field allows spacelike perturbations which may seem to propagate with superluminal velocity. The authors relativized this problem, however, by adding that such perturbations probably “cannot induce acausal effects in the behavior of particles and electromagnetic fields” because these only relate to the conformal factor in the metric (Bekenstein/Milgrom 1984, p. 14). An additional critical point, not only for rAQUAL but for all theories of the original MOND family, stems from their inability to explain the anomalous dynamics in galaxy clusters without assuming some additional unseen matter.

Because of the close relation between integrable Weyl-geometric gravity and JBD theory, Bekenstein’s and Milgrom’s rAQUAL poses the possibility of testing

¹⁴¹Bekenstein, as late as 2004, took it as “evident” that measurements with “clocks and rods” are expressed by the Jordan metric. Moreover, its Levi-Civita connection was taken to govern the free fall of test particles, although the dynamics in the Jordan frame does not satisfy the “usual Einstein equation” (because of explicit terms in the scalar field). The Einstein frame represented for him the “primitive metric” because here the gravitational action reduces to the classical form of the Hilbert term, and the dynamics is given by the Einstein equation (Bekenstein 2004, p. 5f). In their common paper, Milgrom and Bekenstein used the terminology of “dual descriptions” working in “gravitational units” (Einstein frame) respectively “atomic units” (Jordan frame), which sounded a bit like Dirac’s distinction (Bekenstein/Milgrom 1984, p. 14).

¹⁴²In his later review paper Bekenstein qualified this point by stating that only so long as the scalar field “... contributes comparatively little to the energy-momentum tensor, it cannot affect light deflection, which will thus be due to the visible matter alone” (Bekenstein 2004, p. 6). One can read this observation the other way round: If the scalar field carries a considerable contribution to the energy-momentum this influences light deflection.

what happens when it is placed in a scale-invariant framework. In two recent papers, the present author investigated this problem (Scholz 2016a,b). The first paper presents a Weyl-geometrical reformulation of rAQUAL, at least for the so-called “deep MOND” regime and the upper transitional regime. It departs from the view of Bekenstein/Milgrom by showing that with the Weyl-geometric approach observable quantities are most directly expressed in the Einstein gauge. Spacelike components of the Weylian scale connection φ_j , $j = 1, 2, 3$ express additional accelerations beyond those induced by the Riemannian part of the metric (corresponding to Newtonian effects in the weak gravity limit). In the extremely weak gravity regime, two different components for additional accelerations, a_ϕ , a_φ , can be distinguished. The first component is part of the Riemannian acceleration in the Einstein gauge and is due to the energy density of the scalar field ϕ ; the second results from the the Weylian scale connection φ (in the Einstein gauge). In extremely weak static gravitational constellations (i.e., with order of magnitude of Newtonian acceleration a_N close to the MOND acceleration $a_o \approx 1.2 \cdot 10^{-10} \text{ms}^{-2}$), the MOND-like phenomenology is reproduced with results similar to rAQUAL. But here half of the additional acceleration is due to the *scalar field’s energy*, which thus influences the light trajectories.¹⁴³ Whether this suffices for explaining the observed lensing effects remains to be seen.

Moreover, contributions on different length scales to local inhomogeneities of the scalar field’s energy density can add up to produce observable effects that seem to have striking consequences for the dynamics of galaxy clusters. In a heuristic investigation of data from 17(+2) clusters,¹⁴⁴ the author found an encouraging agreement of accelerations predicted by the Weyl-geometric scalar-tensor theory with the corresponding empirical values (Scholz 2016b). This was done on the basis of the observed baryonic masses alone, without assuming additional unseen “dark” matter.¹⁴⁵

Calculations with ordinary MOND, or even its relativistic generalization TeVeS, reduces the need for assuming additional hypothetical dark matter, but one cannot do without it completely. R. Sanders has argued that sterile neutrinos might suffice to fill the gap. In this context it is interesting to note that in the Weyl-geometric framework an *outlandish kinetic term* for the scalar field seems to *suffice for explaining* the otherwise anomalous *dynamics of galaxy clusters*. Of course, there still remain problems. For safeguarding the dynamics on the solar system level the author invented an (ad-hoc) hypothesis postulating that scalar field inhomogeneities are suppressed in regions where the value of at least one sectional curvature of the Riemannian component exceeds a certain threshold (indicating stronger gravitational effects) (Scholz 2016b, p. 6). That would save the dynamics, but it

¹⁴³Cf. the preceding footnote.

¹⁴⁴The data for 2 clusters are outliers, already from the phenomenological point of view.

¹⁴⁵The famous Coma cluster which led Zwicky to introduce the hypothesis of dark matter is among those for which the Weyl-geometric model gives results consistent with most recent empirical data on mass distributions and accelerations.

remains unclear how to account for such a hypothetical suppression. Moreover, the cosmological consequences of this approach are far from clear.¹⁴⁶ In spite of such shortcomings, this model may be of some value for exploring the possibilities of Weyl-geometric scalar fields in the realm of dark matter phenomena.

11.6.4 The Brazilian Approach

Another, and more widely known, challenge to the standard big bang picture, drawing upon Weyl-geometric methods, came from Brazil. Interest in Weyl-geometrical approaches to cosmology have been present in the Brazilian theoretical physics community since the 1990s. The central figure behind this development, *Mário Novello*, acquired his doctorate in 1972 in Geneva under the supervision of J.M. Jauch. Already as a young Ph.D. student, Novello published a paper on Dirac spinors expressed in quaternionic calculus in a Weyl space (Novello 1969).¹⁴⁷ Back in Brazil, working at the *Centro Brasileiro de Pesquisas Físicas* in Rio de Janeiro, he cooperated with many international guests. In 2003 he became the founding director of the *Instituto de Cosmologia Relatividade e Astrofísica (ICRA)*. Due to his influence, Weyl-geometric ideas were introduced in the Brazilian community of theoretical physicists over the course of the 1990s. This activity flourished and formed into a research tradition all its own, the *Brazilian approach* to Weyl-geometric gravity, as I will call it here.

11.6.4.1 A Palatini-Type Path to Integrable Weyl Geometry

In the early 1980s Novello and a co-author from Cologne, *H. Heintzmann*, reflected on the possible consequences for cosmology if one uses models in a slightly more general framework than Riemannian geometry (Novello/Heintzmann 1983). Like other authors before them, they took up a *metric-affine* approach to gravity that utilized a metric g and an independent affine connection Γ . This allows one to define curvature tensors like those in Riemannian geometry, including the scalar curvature R . They began with a gravitational Lagrangian which included a term of the form

$$\mathcal{L}_R = -e^\omega R \sqrt{|g|} \tag{11.110}$$

¹⁴⁶Unpublished calculations indicate scenarios of a cosmic evolution in agreement with many features of standard cosmology. These would support such features as an initial singularity, large parts of the cosmological redshift due to the expansion of spatial folia in Einstein gauge, accelerated “late time” expansion etc.

¹⁴⁷In this paper Novello still thought in terms of Weyl’s first interpretation of the scale connection, the *em dogma* in the terminology above.

with point-dependent function $\omega(x)$,¹⁴⁸ where the metric and affine connection were varied independently according to the so-called *Palatini approach*. They then found that variation with respect to the connection leads to

$$\nabla_\lambda g_{\mu\nu} = -\partial_\lambda \omega g_{\mu\nu} . \quad (11.111)$$

Novello and Heintzmann immediately realized that this relation can be identified with the Weyl-geometrical compatibility condition of equation (11.4) for the integrable scale connection (in our notation)

$$\varphi = \frac{1}{2} d\omega . \quad (11.112)$$

This approach was not without its limitations: it identified the scale gauges in which the coefficient e^ω of the Hilbert term in (11.110) becomes constant with the one in which the scale connection (11.112) vanishes. In terms of our terminology above, this meant there was *no difference* between the *Riemann gauge* and the scalar field – *Einstein gauge*! This structural identification of the two gauges made the Palatini approach to Weyl-geometric gravity amount to a trivial extension of Einstein gravity, if the full scale invariance of the Lagrange density is observed. But such a comparison was not what Novello and Heintzmann had in mind. Referring to Canuto et al. and in the wake of Dirac (cf. Section 11.3.2), they pondered the possibility that atomic clocks and gravitational clocks in different locations might be related by a variable factor $\omega(x)$. If $\omega(x)$ is asymptotically constant, different “Riemannian domains” would arise, possibly connected by “Weyl integrable regions of space”. Moreover, in this setting the “age” of the universe might become “arbitrarily large” (Novello/Heintzmann 1983).

In the years following, Novello developed broad activities in gravitation theory, elementary particle physics, and cosmology; in particular, he was interested in understanding how the initial singularity of standard Riemann-Einstein cosmology could be avoided. In a joint paper with *Edgar Elbaz*, a colleague from France and two members from the Brazilian group, *Jose M. Salim* and *L.A.R. Oliveira*, he and his co-authors proposed an imaginative model for what they called the “creation of the universe” (a clause from the title of the paper) (Novello et al. 1992). Using some Weyl-geometric features and a scalar field ω , they were able to deduce a “cosmic” development from a flat vacuum state (described by Minkowski space) via a contracting phase, “bouncing” at the minimum of a scale function, to an expanding “inflationary” phase. Without going into details of this study, we want to show how and why it became a classical point of reference for the Brazilian tradition in Weyl-geometric methods.¹⁴⁹

¹⁴⁸In this paper $\omega(x)$ was not yet introduced as a scalar field of its own, but via the square of the electromagnetic potential A_μ , i.e., $\omega = \log A_\mu A^\mu$.

¹⁴⁹In many papers of the Brazilian tradition (Novello et al. 1992) is quoted as a starting point: (Salim/Sautú 1996; de Oliveira 1997; Romero et al. 2011, 2012), to cite just a few. Sometimes it

Following (Novello/Heintzmann 1983), Weyl geometry was introduced by the Palatini method of variation (11.111). This led to an integrable Weyl geometry characterized by a scalar function $\omega(x)$, the potential of the scale connection $\varphi = \frac{1}{2}d\omega$. For Novello, Elbaz et al., the above mentioned identification of the Riemann and Einstein gauges did not appear detrimental because their goal was not a modification of Einstein gravity. Rather they set out to model semi-classical quantum “perturbations of the system of measurement units” described by $\delta\omega_\lambda$, such that

$$\delta(\nabla_\lambda g_{\mu\nu}) = (\delta\omega_\lambda)g_{\mu\nu} . \quad (11.113)$$

Perturbations of such a kind are inconsistent with Riemannian geometry but not with Weyl geometry, as the authors noted with references to (Ehlers et al. 1972; Audretsch 1983; Perlick 1991). Like many physicists working in the last third of the 20th century, they thought in terms of a time-evolution of the cosmos, here even in the sense of a *temporal evolution of its geometrical structure*.

They hoped to find “a definite conceptual context ... for the description of such structural transitions” during the cosmic evolution using the Weyl-geometric approach (Novello et al. 1992, p. 650). For this goal, they considered a process governed by the Lagrangian

$$\mathfrak{L}_{vac} = (R + \xi \nabla_\nu \partial^\nu \omega) \sqrt{|g|} \quad (\text{Novello et al. 1992, eqn. (4.2)}), \quad (11.114)$$

where ∇_ν is our notation for the Weyl-geometric derivative and R denotes the Weyl-geometric scalar curvature.¹⁵⁰ From the point of view of Weyl geometry this was a *hybrid approach*: the Lagrange density was not scale invariant, although Weyl-geometric concepts and expressions were used. However, the authors considered this an advantage because they regarded the distinction between “gravitational” units (expressed by a point dependent gravitational “constant”) and atomic units, originally assumed by Dirac, Canuto et al. (see Section 11.3.2.1), as unacceptable. Instead they assumed a broken (active) scale symmetry (Novello et al. 1992, p. 653); thus, they rejected a mere transformation of units in Dicke’s sense, i.e. a passive conception of scale covariance. Understandably, they chose the Einstein gauge for this broken symmetry state. They thus understood (11.114) as an “effective canonical action” of a broken underlying scale symmetric dynamics with some surviving residual Weylian terms. Guided by this physical intuition, the authors avoided a simple reduction to Einstein gravity, which would have become necessary had they assumed full scale invariance.

is even called the “first approach to scalar-tensor theory in WIST” [Weyl integrable space-time] Pucheu et al. (2016).

¹⁵⁰In a side remark the authors reminded that $-2\xi\omega_\lambda\omega^\lambda$ is a variationally equivalent kinetic term because the difference to the kinetic term in (11.114) is a total divergence (Novello et al. 1992, p. 654).

The resulting dynamical equation could be expressed, without loss of content, in Riemannian terms as

$$gR_{\mu\nu} - \frac{gR}{2}g_{\mu\nu} = \lambda^2\omega_\mu\omega_\nu - \frac{\lambda^2}{2}\omega_\alpha\omega^\alpha g_{\mu\nu}, \quad (11.115)$$

with $\lambda^2 = \frac{1}{2}(4\xi - 3)$. This was “equivalent to an Einstein equation in which the WIST¹⁵¹ field ω provides the source of the Riemannian curvature” (Novello et al. 1992, p. 655). As ω was the integral of the scale connection, it had a “purely geometrical origin”. This may have been the reason why they accepted its strange physical properties: negative energy density, positive pressure of the same value (“stiff” matter).

The scalar field equation derived from (11.114) and the Einstein equation (11.115), evaluated for a homogeneous isotropic spacetime, led to a model without initial singularity. In the distant past it looks like a contracting Minkowski space with a non-trivial, “excited” scalar field ω . After a first phase of accelerated contraction, the warp function $a(t)$ reaches a minimum value a_0 , after which an expansionary phase begins. The authors interpreted the first, contracting phase as a vacuum with a geometrical scalar field excitation. Near the minimum they sketched quantum processes of photon and baryon genesis “driven” by the scalar field. Then an expansionary phase follows, ending in a state which, so they argued, could be related to the radiation-dominated phase in the standard model of cosmology. All in all, the calculations were embedded in an imaginative narrative which claimed to solve several pressing problems inherent in the standard picture of the “hot big bang” (no initial singularity, causal horizon and flatness problems, matter anti-matter asymmetry).

11.6.4.2 Cosmological Models with Fluid Matter

Several follow up papers appeared, among them (Salim/Sautú 1996; de Oliveira 1997). In the first, Salim and *S.L. Sautú* added different types of “external fields” representing matter and its interaction with the vacuum Lagrangian (11.114). At first they dealt with an electromagnetic field and an external scalar field, adding matter terms to the Lagrangian, which had a scale-invariant form (Salim/Sautú 1996, eqn. (12)).¹⁵² More important for cosmology, in the next step they adapted to their framework the Lagrangian of a perfect fluid with trajectories that followed a timelike vector field.¹⁵³ Here the hybrid form of their approach with the specified scale gauge was of great advantage because it facilitated the adaptation of the fluid Lagrangian. The authors derived the dynamical equations and constraints in terms

¹⁵¹WIST was (and is) the abbreviation, preferred by the Brazilian authors, for “Weyl integrable spacetime”.

¹⁵²This adds flavour to the hybrid approach mentioned above.

¹⁵³The fluid Lagrangian was taken from (Ray 1972).

of the Weyl-geometric derivative and curvature expressions, in particular taking care for the interaction with the geometrical scalar field ω (Salim/Sautú 1996, eqs. (34)– (40)). They then rewrote the Einstein equation and the scalar field equation in Riemannian terms and then derived the corresponding generalized Raychaudhuri equation for the homogeneous isotropic case (eqn. (47)). Rewriting the coupling constant ξ of (11.114) by $\lambda = \frac{1}{2}(4\xi - 3)$ they concluded that "...depending on the sign of λ , the cosmological solution under consideration can be non-singular and inflationary" (Salim/Sautú 1996, p. 359). This represents a considerable step forward since it generalized the effect observed for the case of the special vacuum solution in (Novello et al. 1992). As the authors rightly concluded:

We have shown that the Weyl integrable geometry can be used in a natural way to geometrize a long-range scalar field. Using a general principle to prescribe the interaction of the geometric scalar field with other physical systems, we can describe in WIST all the classical situations studied by EGR [Einstein gravity, E.S.]. (Salim/Sautú 1996, p. 359)

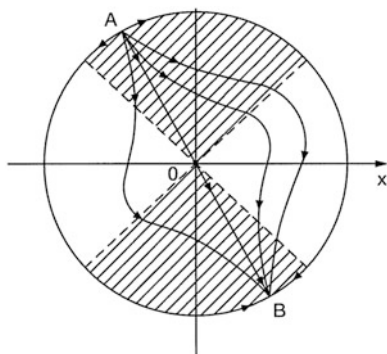
In their next paper, written together with *Henrique P. de Oliveira*, these three authors studied "non-singular inflationary cosmologies in Weyl integrable space-time" (de Oliveira 1997). To the gravitational Lagrangian (11.114) of Novello et al. they added a self-interaction potential of the scalar field $V(\omega)$ and the fluid Lagrangian of (Salim/Sautú 1996). Referring to the same parameter λ as above, they came to the conclusion that for $\lambda > 0$ the Friedmann-like solutions have strong similarities with those of Einstein gravity, while for $\lambda < 0$ interesting "novelties appear in WIST" (p. 2835).

The authors also studied the qualitative behaviour of the modified Friedmann and scalar field equations of their model in the parameter plane (x, y) with

$$x = \frac{\dot{a}}{a}, \quad y = \dot{\omega}. \tag{11.116}$$

For vanishing potential, $V(\omega) = 0$, and for an exponential potential $V(\omega) = V_0 e^{\beta\omega}$ (V_0, β constants) they found that the solutions of the Friedmann equation are generically singularity free, while the solutions with initial or final singularity are unstable (see our Figure 11.1). This was a striking result that led them to

Fig. 11.1 Phase portrait (de Oliveira e.a. 1997, fig 2)



remark: “Depending on the parameter λ , we obtained non-singular models as a general feature. (...) The non-singular behaviour is explained by the violation of the strong energy condition provided by the geometric scalar field” (de Oliveira 1997, p. 2842f). This explanation indicates that the “strong energy condition” was understood in the geometrical sense ($R_{\mu\nu}V^\mu V^\nu > 0$ for any timelike vector field V^μ).¹⁵⁴

11.6.4.3 A Tension Between the Palatini Approach and Scale Invariance

Many more papers on Weyl-geometric gravity were published by the Brazilian group. Some younger researchers also started contributing to the program and published with colleagues from the older generation in different constellations. Among those not yet mentioned were (in alphabetical order): *Tony S. Almeida, F.A.P. Alves, Adriano B. Barreto, Iarley Lobo, José E. Madriz Aguiar, J.B. Fonseca-Neto, Maria L. Pucheu* and *Carlos Romero*. These authors dealt with the relationship of the Palatini variant of Weyl-geometric gravity to Einstein gravity and to JBD theory, and they continued to study the singularity behaviour of cosmological models in their slightly extended framework.

Several of these papers (Romero et al. 2011, 2012) used a Lagrangian of the form:

$$\mathfrak{L} = e^{(1-\frac{n}{2})\omega}(R + 2\Lambda e^{-\omega} + \kappa e^{-\omega} L_m)\sqrt{|g|}, \quad (11.117)$$

where the scalar curvature is to be understood in the metric-affine sense ($R = R(g, \Gamma)$). After a Palatini type variation similar to the transition from (11.110) to (11.111), R turns into the scalar curvature of a Weylian metric given by the pair $(g, \varphi = \frac{1}{2}d\omega)$.¹⁵⁵ The authors emphasize the importance of what appears to them a “new kind of invariance, namely with respect to Weyl transformations” (Romero et al. 2012, 8), however, without strictly maintaining scale invariance as a guiding principle for their investigation. Rewritten in Riemannian terms this Lagrangian acquires the form of a Brans-Dicke Lagrangian with a conformally coupled scalar field. The authors understood this as a “geometrization” of JBD scalar fields in general and, in addition, as an argument for the compatibility of Einstein gravity

¹⁵⁴For the Hawking-Penrose singularity theorem the geometrical energy condition is the crucial point. It has to be distinguished from the *strong energy condition* in the *physical sense* which is given by $T_{\mu\nu} - \frac{1}{2}tr T g_{\mu\nu}V^\mu V^\nu$. The geometrical and physical conditions are equivalent in Einstein gravity; see, e.g., (Curiel 2017, p. 49). Later investigations of the Brazilian school would show that the geometrical energy condition could be violated, while the physical energy condition might still be satisfied (see end of section 5.4.3).

¹⁵⁵(Romero et al. 2011, eq. (7)),(Romero et al. 2012, eqn. (12))

with scale invariance in the sense of Weyl geometry as a wider framework of conventions.¹⁵⁶

Such a generalization of Einstein gravity is too weak to lead to new physical features (see Section 11.6.2.2); it does not even allow one to recuperate the Lagrangian (11.114) so important for the Brazilian tradition. Other papers thus start from a metric-affine generalization of a JBD-type Lagrangian with general coupling coefficient ω , written in the form

$$\mathfrak{L}_{JBD} = e^{-\phi} (R + \omega \partial_\mu \phi \partial^\mu \phi) \sqrt{|g|}, \quad (11.118)$$

with $R = R(g, \Gamma)$ as above. Again the Palatini variation implied the relation (11.111) and, in this way, a motivation for specifying the metric and connection in the sense of integrable Weyl geometry (Pucheu et al. 2014, eqn. (2)). But then the kinetic term, which is taken over without change from the usual Riemannian Brans-Dicke theory, breaks the scale invariance for general ω . Accordingly, these authors use only a restricted Weylian scale transformation. For the transition to the “Einstein frame” they transformed only the quantities $e^{-\phi}$ and R , while leaving the core expression of the kinetic term unaffected (a “field substitution” rather than a gauge transformation), with the result

$$\mathfrak{L} = (\tilde{R} + \omega \partial_\mu \phi \partial^\mu \phi) \sqrt{|g|} \quad (11.119)$$

plus a matter action \mathfrak{L}_m . They were thus led back to the archetypical form (11.114) of the gravitational Lagrangian in the Brazilian tradition.¹⁵⁷

The tension between the Weyl-geometric framework and this mathematical methodology did not pass unnoticed by these authors. But it seems they were prevented from resolving it because of their adherence to the Palatini method of variation and the difficulty of expressing the kinetic term of the scalar field in a scale-invariant form without using Weyl-geometric scale-covariant derivatives (11.89). In the conclusion of one of their papers they wrote “that neither the action nor the field equations of the proposed theory are invariant under Weyl transformations”, admitting that “it would perhaps be desirable, at least from the aesthetic viewpoint, that the whole theory should exhibit Weyl invariance” (Almeida et al. 2014, p. 8). In another paper three of them even gave a twisted explanation for why this allegedly must be so (Pucheu et al. 2014, p. 39). This is surprising, because just two years earlier two of them, supported by Fonseca-Neto, had already noticed that the Lagrangian (11.119) can be scale-transformed to fit the form of the general JBD Lagrangian (11.12). So they were quite close to bringing the Brazilian approach

¹⁵⁶“An important conclusion (...) is that general relativity can perfectly ‘survive’ in a non-Riemannian environment” (Romero et al. 2011). In a personal communication with ES (April 3, 2017) C. Romero added “Philosophically, this could perhaps be interpreted as an illustration of the tenets put forward by Poincaré in his conventionalist ideas”.

¹⁵⁷(Pucheu et al. 2014, eqn. (3.24)), (Almeida et al. 2014, eqn. (16)); compare (11.114) in the light of fn. 150.

into a coherent scale-invariant form;¹⁵⁸ but for some reason these members of the Brazilian group hesitated to transcribe their approach into a scale-invariant mode.

Gravitational Lagrangians of the form (11.119) also played a role in recent qualitative studies of “isotropic cosmologies in Weyl geometry” undertaken by *John Miritzis* from Athens as well as by members of the Brazilian network itself (Miritzis 2004; Pucheu et al. 2016). These qualitative studies deal with cosmological models, either with or without (initial or final) singularities, taking up questions and extending results from the 1990s. A recent paper (Lobo et al. 2015) deals with global singularities in the slightly extended Brazilian framework described above. After a detailed investigation of the Raychaudhuri equation, the authors showed that the geometrical version of the strong energy condition can be violated in the Brazilian approach, while the physical one may be maintained due to contributions of the energy-momentum of the scalar field. This was a sharp result that may well be of wider import. Miritzis, on the other hand, has studied various other aspects of the Brazilian-inspired Weyl-geometric gravity like the “accelerated expansion” of the universe resulting from a scalar field induced fluid (Miritzis 2013a,b).

11.7 Discussion

11.7.1 A Rich History Aside the Mainstream

Our survey dealing with reappearances of Weyl geometry has identified *four different entrance channels* through which central concepts of Weyl’s scale-invariant, but still fully metrical geometry from 1918 were reintroduced into late 20th-century physics. These differed both in motivation and systematics; three of these channels were even opened twice by essentially independent research initiatives with slightly differing systematic ideas (denoted by the word “and” in the following list):

1. Axiomatic foundations of gravity: Ehlers/Pirani/Schild (Section 11.3.1)
2. Scale-co/invariant scalar-tensor theory of gravity: Omote/Utiyama and Dirac (Section 11.3.2)
3. Cartan geometric approach: Bregman and Charap/Tait (Section 11.3.3)
4. de Broglie-Bohm-Madelung (dBMB) approach: Santamato and Shojai/Shojai/Golshani (Section 11.4)

The *first three* were initiated in the short time interval 1971–1974, during which two of the channels were opened twice. Dirac’s motivations for his multiple gauge

¹⁵⁸One only needed to put the JBD Lagrangians in a Weyl-geometric framework. Alternatively, if one wants to start from the Brazilian point of view, one may read the constant coefficient of the Hilbert term in (11.119) as the value of a scale-covariant scalar field χ in (Einstein-) scalar-field gauge, $\chi_o \doteq 1$, and $\partial_\mu\phi\partial^\mu\phi$ as the scalar field gauged expression of the scale-covariant kinetic term $D_\mu\chi D^\mu\chi$ with scale-covariant derivative (11.3).

approach to gravity were quite idiosyncratic and played only a minor role for its wider reception. The broader scenario indicates an intellectual environment that made it appear natural to invoke Weyl's proposal for generalizing Riemannian geometry in a new field-theoretic context. Bjorken scaling had attracted attention in the late 1960s, but already then it was known to be only approximatively valid, so it could not have been a major driving force.¹⁵⁹ On the other hand, the field structures of elementary particle physics were just acquiring the form and status of a new, gauge-theoretic standard model, due to the renormalizability results of 't Hooft and Veltman (1972) and the experimental detection of quark binding states, called "J- Ψ " (1974). Their Lagrangians were basically (globally) scale invariant in Minkowski space, with only the mass term of the hypothetical Higgs field as a scale breaking term. This context may have strongly motivated researches which explored possible connections to gravity in an enlarged scale-covariant framework. From this wider perspective a new look at Weyl-geometric generalizations of Einstein gravity must have appeared promising. In this respect it was important that the Jordan-Brans-Dicke research program of scalar-tensor theories had already shown a decade earlier how one could model gravity without taking recourse to a quadratic curvature term, thereby suggesting that it would be natural to do so also in a Weyl-geometric setting. This brought the Weylian approach much closer to Einstein gravity than the quadratic gravity theories studied since the time of Weyl, which was doubly important given the fact that during this time Einstein gravity passed through a phase of new empirical confirmations.¹⁶⁰

The *fourth channel* was anchored in the completely different intellectual context of the de Broglie-Bohm program for reconsidering the foundations of quantum mechanics. Its two research lines began a decade later than the first three (Santamato), or even two decades later (Shojai/Shojai/Golshani), the "Tehran approach". These two lines differed from each other more strikingly than did the respective double starts in items 2 and 3. In the 1990s the Bohmian approaches entered a latency phase (Santamato) or were just heading towards a new beginning, still developed within a JBD framework (Iranian approach). The authors in the latter group started to use Weyl-geometric concepts explicitly only after the turn into the new millennium.

All in all, the time until roughly 2000 was a *first phase of exploration* for all these approaches. For several years the immediate contributors to the Dirac line explored the astrophysical consequences of Dirac's distinction between an "atomic gauge" and an "Einstein gauge" (Bouvier, Maeder, Canuto et al.) or they refined and extended the theory (Rosen). At first they stuck with Dirac's interpretation of the scale connection as an electromagnetic potential, the *em dogma*, but by the 1980s this allegiance to Dirac's ideas faded away. Those who continued to appeal to his approach, like Rosen and Israelit, enriched the perspective by considering the scale connection as a representative of a Proca-like massive gauge field or, like Smolin,

¹⁵⁹Only for the authors of (Hehl et al. 1988c) did this appear differently, see Section 11.3.3.

¹⁶⁰Cf. C. Will's contribution to this volume.

placed it in a different context. This boiled down to a merging of the modified Dirac line with the research stemming from the initiative of Omote/Utiyama (Hayashi, Kugo et al.), which intensified with attempts to bring Weyl's scale geometry in contact with the field content of the rising standard model (see below).

On the other hand, the later researches of Rosen and Israelit explored a vast terrain of theoretical possibilities, many of them quite speculative. They inquired as to whether the energy-momentum of a Weyl-geometric gravitational scalar field or a hypothetical "Weylon gas" might contribute to dark matter phenomena and/or to the accelerated expansion diagnosed in the usual Riemannian approach to gravity. But these studies were on a quite general level and remained mainly without closer links to astrophysical or astronomical observations (Section 11.6.2.1).

The *Cartan-Weyl-geometric approach* was soon relegated to a very special case within the broader Cartan geometric metric-affine theories (Hehl et al.) or else it was studied in relation to Kaluza-Klein theory.¹⁶¹ As the latter are not included in this survey, these lines of investigation fall outside the range of our panorama. The foundational studies of Ehlers/Pirani/Schild, on the other hand, found a broad and continued reception and development in the philosophy of physics and remained a point of orientation for foundational studies of gravity.

For some authors (Englert, Smolin, Cheng, later Drechsler, Tann) the *rise of the standard model* motivated their attempts to connect Weyl-geometric gravity – or at least scale-covariant gravity in the case of Englert et al. – with standard model fields, in particular the Higgs field. Cheng's seminal paper of 1988 was the first relatively detailed account of the electroweak sector of the SM assimilated into a Weyl-geometric context, although only on the pseudo-classical level of the theory (Section 11.5.1).¹⁶² Nearly a decade later, Drechsler and Tann derived much of the electroweak structure in their own development of Weyl geometry,¹⁶³ but with the unusual idea of considering the Higgs field as a part of the gravitational structure.

In the *new century* this idiosyncratic idea was superseded by the studies of Nishina and Rajpoot, who continued the research opened up by Cheng while staying closer to the mainstream expectations of a massive Higgs field, which at that time was still hypothetical (Section 11.5.3.1). Here we finally find explicit references to the papers of Dirac and those in the Utiyama research tradition, indicating a merging of the lines mentioned above. With the empirical detection of the Higgs quantum excitation ("particle"), this line gained momentum as the most realistic among the Weyl-geometric approaches to SM fields. But the question as to how scale symmetry is related to the quantum level still remains open. Ohanian's attempt to show that scale symmetry is "spontaneously broken" near the Planck scale, and thus leads

¹⁶¹E.g. (Drechsler/Hartley 1994).

¹⁶²Studies of QFT on Weylian manifolds, comparable to the corresponding researches for Lorentzian manifolds, as discussed in R. Wald's contribution to this volume, are still a desideratum.

¹⁶³These two authors referred neither to the Dirac tradition in Weyl-geometric gravity nor to Utiyama's; their Weyl-geometric starting point was "self-made" (Drechsler/Hartley 1994) aside from Weyl's original papers.

back to Einstein gravity, may be considered as a nice toy model (Section 11.5.4.2), but the investigations of Codello et al. indicate that the last word here has not yet been spoken (Section 11.5.4.1)

With regard to astrophysics and cosmology, the first exploratory phase of investigations was superseded by the Brazilian research tradition of Weyl-geometric gravity, initiated by the work of Novello et al. Although this research line has been confined to a geometrically “hybrid” approach – which would imply a dynamically inert scalar field if the Weyl-geometric scaling symmetry were to be taken seriously – this group of authors followed the physical intuition of their “founding father”. Their viewpoint thus goes back to the paper that launched this tradition, (Novello et al. 1992), which assumed an effective action of a broken underlying scale symmetry with some surviving residual Weylian terms (Section 11.6.4). The advantage of such an approach is that it allows one to investigate concrete cosmological models that give some impression of the possibilities a Weyl-geometric extension might offer compared with the present-day Riemann-Einstein cosmological models. From a different side, a bridge to the family of MOND-like theories of dark matter has also been established, thereby widening the research horizon for this Weyl-geometric extension of gravity theories even further (Section 11.6.3).

All in all the scale-covariant and often explicitly Weyl-geometric approaches to gravity, elementary particle fields, foundations of quantum mechanics, astrophysics and cosmology have developed a rich array of models since the 1970s. In many cases lesser known scientists contributed to this research, but the Weylian framework has occasionally attracted the attention of internationally renowned physicists. Although the Weyl-geometric perspective has remained up to now a side-stream in all of these fields, it may well offer interesting challenges and possibilities for the future.

11.7.2 ... and an Open Research Horizon

Our panorama has shown a variety of approaches which do yet not form a coherent research program. The Bohmian research lines, e.g., still stand apart from the other approaches, although some formal connections to the scalar fields of Dirac/Omote/Utiyama type have been established in the later phase. It is not clear, however, whether the “Tehran” perspective stands on solid ground, and if so whether it can be integrated with the “Italian” approach into a consistent common picture. Other threads within the whole field indicate perspectives that may potentially reinforce one another. The following remarks pertaining to this “open research horizon” are necessarily subjective, but they may be useful for a general orientation.

Although cosmology has increased its observational basis in the past few decades tremendously, it continues to call for alternative approaches to deal with its many conundrums. Several other contributions to this volume (H. Kragh, C. Smeenk, M. Gaberdiel) discuss some of these alternatives. The Weyl-geometric approach plays a role in this challenge, although for the time being only a minor one. The work of the Brazilian group has made the most concrete contribution to this subject,

but it is still hampered by the constraints resulting from reliance on the Palatini approach to variation (Section 11.6.4). Moreover, they have not yet explored the full consequences of the rescaling freedom in Weyl-geometric Robertson-Walker models, including the possibility that part of the cosmological redshift may be due to the scale connection rather than to a “real” expansion (Section 11.6.2.2). Such a turn towards a more field-theoretic explanation of the cosmological redshift would open new vistas for the geometry of cosmological model building, for which the Weyl-geometric approach is clearly well suited.

A recent paper by (Codello et al., Section 11.5.4) takes first steps towards a field quantization scheme in a Weyl-geometric environment that preserves scale symmetry at the quantum level. If this quantization procedure, or other ones that preserve the scale symmetry, can be extended to the complete set of standard model fields plus the Weyl-geometric scale connection and the gravitational scalar field, we may arrive at a *modest* integration of gravity and the SM, in which only the scale degree of freedom of the metric is quantized. Bars/Steinhardt/Turok have already argued that a theory with scale symmetry at the quantum level may lead to a cancelling of the quadratically divergent terms in the radiative corrections to the Higgs mass.¹⁶⁴ Although this is still an unproven expectation, it nevertheless should be regarded as a highly interesting observation that goes to the hard core of the *naturalness problem* in present elementary particle physics. Taken together with the long standing speculations regarding the scalar field and/or the “Weylon” (scale connection) field as candidates for dark matter (Sections 11.3.2.2 and 11.6.2.1), the Weyl-geometric approach would seem to offer chances for attacking the naturalness problem of the SM and the dark matter problem jointly, essentially by extending the underlying automorphism group of gravity and quantum field theory.

This complex of expectations has fed much of the research dynamics of the supersymmetry program, so we seem to be approaching a similar thematic complex here, though in a more modest form. We also have seen that a classical, “effective” view of the gravitational field can lead to MOND-like phenomenology if unusual kinematical terms for the scalar field are also taken into account (Section 11.6.3). We can thus look forward with interest and curiosity to see where future research will eventually lead.

Acknowledgements This paper owes its existence to *David Rowe’s* initiative in several respects. He encouraged me to present heterodox ideas on Weyl-geometric methods in cosmology at the conference and invited me to rethink the case after a cool reception of the talk by the other participants. That gave me the chance to place my views in the wider range of recent attempts to use Weyl-geometric methods in physics. After an interruption of several years, an earlier first draft of this paper (Scholz 2011b) had to be rewritten completely for the final version of this book. The new version overlaps nicely with the wider ambit of the investigations of the interdisciplinary group *Epistemology of the LHC* with center at Wuppertal and supported generously by the DFG/FWF. This group offers the chance for a close communication between historians and philosophers of science and colleagues from the elementary particle community. H. Cheng, F. Hehl, J. Miritzis, C. Romero, D. Rowe, A. Trautman, S. Walter gave helpful hints for the final version of the paper.

¹⁶⁴(Bars et al. 2014, p. 2)

References

- Abraham, R., & Marsden, J. E. (1978). *Foundations of classical mechanics*. Redwood City: Addison Wesley. 5-th revised ed. 1985.
- Adler, S. (1982). Einstein gravity as a symmetry-breaking effect in quantum field theory. *Reviews of Modern Physics*, 54, 729–766.
- Almeida, T. S., Pucheu, M. L., Romero, C., & Formiga, J. B. (2014). From Brans-Dicke gravity to a geometrical scalar-tensor theory. *Physical Review D*, 89, 064047. arXiv:1311.5459.
- Audretsch, J. (1983). Riemannian structure of space-time as a consequence of quantum mechanics. *Physical Review D*, 27, 2872–2884.
- Audretsch, J., Gähler, F., & Straumann, N. (1984). Wave fields in Weyl spaces and conditions for the existence of a preferred pseudo-riemannian structure. *Communications in Mathematical Physics*, 95, 41–51.
- Audretsch, J., Hehl, F. W., & Lämmerzahl, C. (1992). Matter wave interferometry and why quantum objects are fundamental for establishing a gravitational theory. In J. Ehlers & G. Schaefler (Eds.), *Relativistic gravity research with emphasis on experiments and observations* (vol. 410, pp. 369–407). *Lecture notes in physics*. Berlin: Springer.
- Audretsch, J., & Lämmerzahl, C. (1988). Constructive axiomatic approach to spacetime torsion. *Classical and Quantum Gravity*, 5, 1285–1295.
- Audretsch, J., & Lämmerzahl, C. (1991). Establishing the Riemannian structure of space-time by means of light rays and free matter waves. *Journal of Mathematical Physics*, 32, 2099–2105.
- Audretsch, J., & Lämmerzahl, C. (1994). A new constructive axiomatic scheme for the geometry of space-time. In U. Majer & H.-J. Schmidt (Eds.), *Semantical Aspects of Spacetime Theories* (pp. 21–40). Mannheim: BI-Verlag.
- Bacciagaluppi, G. (2005). A conceptual introduction to Nelson's mechanics. In R. Buccheri, A. Elitzur, & M. Saniga (Eds.), *Endophysics, time, quantum and the subjective* (pp. 367–388). Singapore: World Scientific. Revised postprint in philsci-archive.pitt.edu/8853/1/Nelson-revised.pdf.
- Bacciagaluppi, G., & Valentini, A. (2009). *Quantum theory at the crossroads. Reconsidering the 1927 solvay conference*. Cambridge: Cambridge University Press.
- Bars, I., Steinhardt, P., & Turok, N. (2014). Local conformal symmetry in physics and cosmology. *Physical Review D*, 89, 043515. arXiv:1307.1848.
- Bekenstein, J. (2004). Relativistic gravitation theory for the modified Newtonian dynamics paradigm. *Physical Review D*, 70, 083509.
- Bekenstein, J., & Milgrom, M. (1984). Does the missing mass problem signal the breakdown of Newtonian gravity? *Astrophysical Journal*, 286, 7–14.
- Bergmann, P. G. (1942). *Introduction to the theory of relativity*. Englewood Cliffs: Prentice Hall. Reprint New York: Dover 1976.
- Blagojević, M. (2002). *Gravitation and gauge symmetries*. Bristol/Philadelphia: Institute of Physics Publishing.
- Blagojević, M., & Hehl, F. W. (2013). *Gauge theories of gravitation. A reader with commentaries*. London: Imperial College Press.
- Bohm, D. (1952a). A suggested interpretation of the quantum theory in terms of 'hidden variables' I. *Physical Review*, 85(1), 166–179.
- Bohm, D. (1952b). A suggested interpretation of the quantum theory in terms of 'hidden variables' II. *Physical Review*, 85(2), 180–193.
- Borrelli, A. (2015). The story of the Higgs boson: the origin of mass in early particle physics. *European Physical Journal H*, 40(1), 1–52.
- Bouvier, P., & Maeder, A. (1977). Consistency of Weyl's geometry as a framework for gravitation. *Astrophysics and Space Science*, 54, 497–508.
- Brans, C. (1961). Mach's principle and a varying gravitational constant. PhD thesis, Physics Department, Princeton University

- Brans, C. (1999). Gravity and the tenacious scalar field. In A. Harvey (Ed.), *On Einstein's path. Essays in honor of engelbert schücking* (pp. 121–138). Berlin: Springer. arXiv:gr-qc/9705069.
- Brans, C. (2005). The roots of scalar-tensor theories: an approximate history. In *International workshop on gravitation and cosmology*, Contributions to the Cuba Workshop, Santa Clara 2004. arXiv:gr-qc/0506063.
- Brans, C. (2014). Jordan-Brans-Dicke theory. *Scholarpedia*, 9(4), 31358.
- Brans, C., & Dicke, R. H. (1961). Mach's principle and a relativistic theory of gravitation. *Physical Review*, 124, 925–935.
- Bregman, A. (1973). Weyl transformations and Poincaré gauge invariance. *Progress of Theoretical Physics*, 49, 667–6992.
- Cai, R.-G., & Wei, H. (2007). Cheng-Weyl vector field and its cosmological application. *Journal of Cosmology and Astroparticle Physics*, 0709, 015. arXiv:astro-ph/0607064.
- Calderbank, D., & Pedersen, H. (1998). Einstein-Weyl geometry. *Advances in Mathematics*, 97, 74–109.
- Callan, C., Coleman, S., & Jackiw, R. (1970). A new improved energy-momentum tensor. *Annals of Physics*, 59, 42–73.
- Canuto, V., Adams, P. J., Hsieh S.-H., & Tsiang, E. (1977). Scale covariant theory of gravitation and astrophysical application. *Physical Review D*, 16, 1643–1663.
- Canuto, V., & Goldman, I. (1983). Astrophysical consequences of a violation of the strong equivalence principle. *Nature*, 304, 311–314.
- Capozziello, S., & Faraoni, V. (2011). *Beyond Einstein gravity. A survey of gravitational theories for cosmology and astrophysics*. Dordrecht: Springer.
- Carroll, R. (2004). Gravity and the quantum potential. arXiv:gr-qc/0406004.
- Castro, C. (1992). On Weyl geometry, random processes, and geometric quantum mechanics. *Foundations of Physics*, 22, 569–615.
- Castro, C. (2007). On dark energy, Weyl's geometry, different derivations of the vacuum energy density and the Pioneer anomaly. *Foundations of Physics*, 37, 366–409.
- Castro, C. (2009). The cosmological constant and Pioneer anomaly from Weyl geometry and Mach's principle. *Physics Letters B*, 675, 226–230.
- Charap, J. M., & Tait, W. (1974). A gauge theory of the Weyl group. *Proceedings Royal Society London A*, 340, 249–262.
- Chen, P., & Kleinert, H. (2016). Deficiencies of Bohm trajectories in view of basic quantum principles. *Electronic Journal of Theoretical Physics*, 13(35), 1–12.
- Cheng, H. (1988). Possible existence of Weyl's vector meson. *Physical Review Letters*, 61, 2182–2184.
- Clifton, T., Ferreira, P., Padilla, A., & Skordis, C. (2012). Modified gravity and cosmology. *Physics Reports*, 513, 1–189. arXiv:1106.2476.
- Codello, A., D'Orodoico, G., Pagani, C., & Percacci, R. (2013). The renormalization group and Weyl invariance. *Classical and Quantum Gravity*, 30, 115015. arXiv:1210.3284.
- Coleman, R., & Korté, H. (1984). Constraints on the nature of inertial motion arising from the universality of free fall and the conformal causal structure of spacetime. *Journal of Mathematical Physics*, 25, 3513–3526.
- Coleman, S., & Weinberg, E. (1973). Radiative corrections as the origin of spontaneous symmetry breaking. *Physical Review D*, 7, 1888–1910.
- Cotsakis, S., & Miritzis, J. (1999). Variational and conformal structure of nonlinear metric-connection gravitational Lagrangians. *Journal of Mathematical Physics*, 40(6), 3063.
- Curiel, E. (2017). A primer on energy conditions. In (Lehmkuhl et al. 2017, pp. 43–104).
- Dahia, F., Gomez, A. T., & Romero, C. (2008). On the embedding of space-time in five-dimensional Weyl spaces. *Journal of Mathematical Physics*, 49, 102501. arXiv:0711.2754.
- Darrigol, O. (2014). *Physics and necessity: Rationalist pursuits from the cartesian past to the quantum present*. Oxford: Oxford University Press.
- de Broglie, L. (1960). *Non-linear wave mechanics. A causal interpretation*. Amsterdam: Elsevier. Translated by A.J. Knodel and J.C. Miller.

- De Martini, F., & Santamato, E. (2014a). Interpretation of quantum-nonlocality by conformal geometrodynamics. *International Journal of Theoretical Physics*, *53*, 3308–3322. arXiv:1203:0033.
- De Martini, F., & Santamato, E. (2014b). Nonlocality, no-signalling, and Bell's theorem investigated by Weyl conformal differential geometry. *Physica Scripta*, *2014*, T163. arXiv:1406.2970.
- De Martini, F., & Santamato, E. (2014c). The intrinsic helicity of elementary particles and the spin-statistic connection. *International Journal of Quantum Information*, *12*, 1560004.
- De Martini, F., & Santamato, E. (2015). Proof of the spin-statistics theorem. *Foundations of Physics*, *45*(7), 858–873.
- De Martini, F., & Santamato, E. (2016). Proof of the spin-statistics theorem in the relativistic regime by Weyl's conformal quantum mechanics. *International Journal of Quantum Information*, *14*(04), 1640011.
- de Oliveira, H. P., Salim, J. M., & Sautú, S. L. (1997). Non-singular inflationary cosmologies in Weyl integrable spacetime. *Classical and Quantum Gravity*, *14*(10), 2833–2843.
- Deser, S. (1970). Scale invariance and gravitational coupling. *Annals of Physics*, *59*, 248–253.
- Dicke, R. H. (1962). Mach's principle and invariance under transformations of units. *Physical Review*, *125*, 2163–2167.
- Dirac, P. A. M. (1973). Long range forces and broken symmetries. *Proceedings Royal Society London A*, *333*, 403–418.
- Dirac, P. A. M. (1974). Cosmological models and the large number hypothesis. *Proceedings Royal Society London A*, *338*, 439–446.
- Drechsler, W. (1999). Mass generation by Weyl symmetry breaking. *Foundations of Physics*, *29*, 1327–1369.
- Drechsler, W., & Hartley, D. (1994). The role of the internal metric in generalized Kaluza-Klein theories. *Journal of Mathematical Physics*, *35*, 3571–3585.
- Drechsler, W., & Mayer, M. E. (1977). *Fibre bundle techniques in gauge theories. Lectures in mathematical physics at the University of Austin* (vol. 67) Lecture notes in physics. Berlin: Springer.
- Drechsler, W., & Tann, H. (1999). Broken Weyl invariance and the origin of mass. *Foundations of Physics*, *29*(7), 1023–1064. arXiv:gr-qc/98020.
- Dürr, D., Goldstein, S., Tumulka, R., & Zanghi N. (2009). Bohmian mechanics. In D. Greenberger, K. Hentschel, & F. Weinert (Eds.), *Compendium of quantum physics* (pp. 47–55). Berlin: Springer.
- Eddington, A. S. (1923). *The mathematical theory of relativity*. Cambridge: Cambridge University Press.
- Ehlers, J., Pirani, F., & Schild, A. (1972). The geometry of free fall and light propagation. In L. O'Raifeartaigh (Ed.), *General relativity, papers in honour of J.L. Synge* (pp. 63–84). Oxford: Clarendon Press.
- Einstein, A. (1916). Die Grundlagen der allgemeinen Relativitätstheorie. *Mathematische Annalen*, *49*, 769–822.
- Einstein, A. (1949). *Autobiographical notes* (vol. 7) The library of living philosophers. La Salle, IL: Open Court.
- Einstein, A. (1998). *The collected papers of Albert Einstein. Volume 8. The Berlin years: Correspondence, 1914–1918, Part B: 1918*. R. Schulmann, A. J. Kox, M. Janssen, J. Illy, & K. von Meyenn (Eds.). Princeton: Princeton University Press.
- Englert, F., Gunzig, E., Truffin, C., & Windey, P. (1975). Conformal invariant relativity with dynamical symmetry breakdown. *Physics Letters*, *57 B*, 73–76.
- Englert, F., & Truffin, C. (1976). Conformal invariance in quantum gravity. *Nuclear Physics B*, *117*, 407–432.
- Faraoni, V., & Nadeau, S. (2007). (Pseudo)issue of the conformal frame revisited. *Physical Review D*, *75*(2), 023501.
- Flato, M., & Raçka, R. (1988). A possible gravitational origin of the Higgs field in the standard model. *Physics Letters B*, *208*, 110–114. Preprint, SISSA (Scuola Internazionale Superiore di Studi Avanzate), Trieste, 1987 107/87/EP.

- Flato, M., & Simon, J. (1972). Wightman formulation for the quantization of the gravitational field. *Physical Review D*, 5, 332–341.
- Folland, G. B. (1970). Weyl manifolds. *Journal of Differential Geometry*, 4, 145–153.
- Foot, R., & Kobakhidze, A. (2013). Electroweak scale invariant models with small cosmological constant. *International Journal of Modern Physics A*, 30(21), 1550126. arXiv:0709.2750.
- Foot, R., Kobakhidze, A., McDonald, K., & Volkas, R. R. (2007a). Neutrino mass in radiatively-broken scale-invariant models. *Physical Review D*, 76, 075014. arXiv:0706.1829.
- Foot, R., Kobakhidze, A., McDonald, K., & Volkas, R. R. (2007b). A solution to the hierarchy problem from an almost decoupled hidden sector within a classically scale invariant theory. *Physical Review D*, 77, 035006. arXiv:0709.2750.
- Franklin, A. (2017). The missing piece of the puzzle: The discovery of the Higgs boson. *Synthese*, 194(2), 259–274. <https://doi.org/10.1007/s11229-014-0550-y>.
- Fujii, Y., & Maeda, K.-C. (2003). *The scalar-tensor theory of gravitation*. Cambridge: Cambridge University Press.
- Gauduchon, P. (1995). La 1-forme de torsion d'une variété hermitienne compacte. *Journal für die reine und angewandte Mathematik*, 469, 1–50.
- Gilkey, P., Nikčević, S., & Simon, U. (2011). Geometric realizations, curvature decompositions, and Weyl manifolds. *Journal of Geometry and Physics*, 61, 270–275. arXiv:1002.5027.
- Goenner, H. (2004). On the history of unified field theories. *Living Reviews in Relativity*, 7, 2. <http://relativity.livingreviews.org/Articles/lrr-2004-2>.
- Goenner, H. (2012). Some remarks on the genesis of scalar-tensor theories. *General Relativity and Gravity*, 44(8), 2077–2097. arXiv: 1204.3455.
- Gray, J. (Ed.) (1999). *The symbolic universe: Geometry and physics 1890–1930*. Oxford: Oxford University Press.
- Hall, B. C. (2001). *Quantum theory for mathematicians*. Berlin: Springer.
- Hayashi, K., & Kugo, T. (1979). Remarks on Weyl's gauge field. *Progress of Theoretical Physics*, 61, 334–346.
- Hehl, F. W. (1970). Spin und Torsion in der allgemeinen Relativitätstheorie oder die Riemann-Cartansche Geometrie der Welt. *Habilitationsschrift*. Technische Universität Clausthal: Mimeograph.
- Hehl, F. W. (2017). Gauge theory of gravity and spacetime. In (Lehmkuhl et al. 2017, 145–170).
- Hehl, F. W., Kerlick, G. D., & von der Heyde, P. (1976a). On a new metric affine theory of gravitation. *Physics Letters B*, 63 (4), 443–448.
- Hehl, F. W., McCrea, J. D., & Kopczyński, W. (1988a). The Weyl group and its currents. *Physics Letters A*, 128, 313–318.
- Hehl, F. W., McCrea, J. D., & Mielke, E. (1988b). Skaleninvarianz und Raumzeit-Struktur. In B. Geyer, H. Herwig, & H. Rechenberg (Eds.), *Werner Heisenberg. Physiker und Philosoph* (pp. 299–306). Berlin: Spektrum.
- Hehl, F. W., McCrea, J. D., Mielke, E., & Ne'eman, Y. (1989). Progress in metric-affine theories of gravity with local scale invariance. *Foundations of Physics*, 19, 1075–1100.
- Hehl, F. W., McCrea, J. D., Mielke, E., & Ne'eman, Y. (1995). Metric-affine gauge theory of gravity: Field equations, noether identities, world spinors, and breaking of dilation invariance. *Physics Reports*, 258, 1–171.
- Hehl, F. W., Mielke, E., & Tresguerres, R. (1988c). Weyl spacetimes, the dilation current, and creation of gravitating mass by symmetry breaking. In W. Deppert & K. Hübner (Eds.), *Exact sciences and their philosophical foundations; exakte wissenschaften und ihre philosophische grundlegung* (pp. 241–310). Frankfurt: Peter Lang.
- Hehl, F. W., Puntigam, R., & Tsantilis, E. (1996). A quadratic curvature Lagrangian of Pawłowski and Raczka: A finger exercise with MathTensor. In F. W. Hehl, R. Puntigam, & H. Ruder (Eds.), *Relativity and scientific computing*. . . Berlin: Springer. [gr-qc/9601002].
- Hehl, F. W., von der Heyde, P., Kerlick, G. D., & Nester, J. M. (1976b). General relativity with spin and torsion: Foundations and prospects. *Reviews of Modern Physics*, 48, 393–416.
- Higa, T. (1993). Weyl manifolds and Einstein-Weyl manifolds. *Commentarii Mathematici Sancti Pauli*, 42, 143–160.

- Israelit, M. (1996). Conformally coupled dark matter. *Astrophysics and Space Science*, 240(1), 331–344. arXiv:gr-qc/9608035.
- Israelit, M. (1999a). Matter creation by geometry in an integrable Weyl–Dirac theory. *Foundations of Physics*, 29(8), 1303–1322.
- Israelit, M. (1999b). *The Weyl-Dirac theory and our universe*. New York: Nova Science.
- Israelit, M. (2002a). Primary matter creation in a Weyl-Dirac cosmological model. *Foundation of Physics*, 32, 295–321.
- Israelit, M. (2002b). Quintessence and dark matter created by Weyl-Dirac geometry. *Foundation of Physics*, 32, 945–961.
- Israelit, M. (2010). A Weyl-Dirac cosmological model with DM and DE. *General Relativity and Gravitation*, 43, 751–775. arXiv:1008.0767.
- Israelit, M. (2012). Nowadays cosmology with the Weyl-Dirac approach. Preprint arXiv:1212.2208. Slightly changed version of Israelit (2010).
- Israelit, M., & Rosen, N. (1992). Weyl-Dirac geometry and dark matter. *Foundations of Physics*, 22, 555–568.
- Israelit, M., & Rosen, N. (1993). Weylian dark matter and cosmology. *Foundations of Physics*, 24, 901–915.
- Israelit, M., & Rosen, N. (1995). Cosmic dark matter and Dirac gauge function. *Foundations of Physics*, 25, 763–777.
- Jordan, P. (1952). *Schwerkraft und weltall*. Braunschweig: Vieweg. 2nd revised edition 1955.
- Kaiser, D. (2006). Whose mass is it anyway? Particle cosmology and the objects of a theory. *Social Studies of Science*, 36(4), 533–564.
- Kaiser, D. (2007). When fields collide. *Scientific American*, pp. 62–69.
- Karaca, K. (2013). The construction of the Higgs mechanism and the emergence of the electroweak theory. *Studies in History and Philosophy of Modern Physics*, 44, 1–16.
- Kasuya, M. (1975). On the gauge theory in the Einstein–Cartan–Weyl space-time. *Nuovo Cimento B*, 28(1), 127–137.
- Kibble, T. (1961). Lorentz invariance and the gravitational field. *Journal for Mathematical Physics*, 2, 212–221. In (Blagojević/Hehl 2013, chap. 4).
- Kleinert, H. (2008). *Multivalued fields in condensed matter, electromagnetism, and gravitation*. Singapore: World Scientific.
- Kosmann-Schwarzbach, Y. (2011). *The noether theorems. invariance and conservation laws in the twentieth century*. Berlin: Springer.
- Kostant, B. (1970). *Quantization and unitary representations. I. Prequantisation* (vol. 170) Lecture notes in mathematics. Berlin: Springer.
- Kragh, H. (1999). *Quantum generations: A history of physics in the twentieth century*. Princeton: Princeton University Press.
- Kragh, H. (2006). Cosmologies with varying speed of light: A historical perspective. *Studies in History and Philosophy of Modern Physics*, 37, 726–737.
- Kragh, H. (2009). Continual fascination: The oscillating universe in modern cosmology. *Science in Context*, 22(4), 587–612.
- Kragh, H. (2016). *Varying gravity. Dirac’s legacy in cosmology and geophysics. Science networks*. Heidelberg: Springer-Birkhäuser.
- Lämmerzahl, C. (1990). The geometry of matter fields. In V. de Sabbata & J. Audretsch (Eds.), *Quantum mechanics in curved spacetime* (pp. 23–48). Berlin: Springer.
- Lehmkuhl, D. (2014). Why Einstein did not believe that general relativity geometrizes gravity. *Studies in History and Philosophy of Modern Physics*, 46B, 316–326. <http://philsci-archiv.pitt.edu/9825/>.
- Lehmkuhl, D., Schieman, G., & Scholz, E. (Eds.) (2017). *Towards a theory of spacetime theories. Einstein studies*. Berlin/Basel: Springer/Birkhäuser.
- Lobo, I. I., Barreto, A. B., & Romero, C. (2015). Space-time singularities in Weyl manifolds. *European Physics Journal C*, 75 (9), 448. arXiv:1506.02180.
- Madelung, E. (1926). Quantentheorie in hydrodynamischer Form. *Zeitschrift für Physik*, 40(3–4), 322–326.

- Maeder, A. (1978a). Metrical connection in space-time, Newton's and Hubble's law. *Astronomy and Astrophysics*, 65, 337–343.
- Maeder, A. (1978b). Cosmology II: Metrical connection and clusters of galaxies. *Astronomy and Astrophysics*, 67, 81–86.
- Mannheim, P. (1994). Open questions in classical gravity. *Foundations of Physics*, 224, 487–511.
- Mannheim, P. (2000). Attractive and repulsive gravity. *Foundations of Physics*, 22, 709–746.
- Mannheim, P. (2012). Making the case for conformal gravity. *Foundations of Physics*, 42. arXiv:1101.2186.
- Mannheim, P., & Kazanas, D. (1989). Exact vacuum solution to conformal Weyl gravity and galactic rotation curves. *Astrophysical Journal*, 342, 635–638.
- Meissner, K., & Nicolai, H. (2009). Conformal symmetry and the standard model. *Physics Letters B*, 648, 312–317. arXiv:hep-th/0612165.
- Miritzis, J. (2004). Isotropic cosmologies in Weyl geometry. *Classical and Quantum Gravity*, 21, 3043–3056. arXiv:gr-qc/0402039.
- Miritzis, J. (2013a). Energy exchange in Weyl geometry. In *Proceedings of the Greek Relativity Meeting NEB15, June 2012, Chania, Greece*. Journal of physics: Conference series. arXiv:1301.5402.
- Miritzis, J. (2013b). Acceleration in Weyl integrable spacetime. *International Journal of Modern Physics D*, 22(5), 1350019. arXiv:1301.5696.
- Myrvold, W. (2003). On some early objections to Bohm's theory. *International Studies in the Philosophy of Science*, 17(1), 7–24.
- Narlikar, J., & Padmanabhan, T. (1983). Quantum cosmology via path integrals. *Physics Reports*, 100, 151–200.
- Nicolic, H. (2005). Relativistic quantum mechanics and the Bohmian interpretation. *Foundations of Physics Letters*, 18(6), 549–561.
- Nieh, H.-T. (1982). A spontaneously broken conformal gauge theory of gravitation. *Physics Letters A*, 88, 388–390.
- Nishino, H., & Rajpoot, S. (2004). Broken scale invariance in the standard model. arXiv:hep-th/0403039.
- Nishino, H., & Rajpoot, S. (2007). Broken scale invariance in the standard mode. *AIP Conference Proceedings*, 881, 82–93. arXiv:0805.0613 (with different title).
- Nishino, H., & Rajpoot, S. (2009). Implication of compensator field and local scale invariance in the standard model. *Physical Review D*, 79, 125025. arXiv:0906.4778.
- Nishino, H., & Rajpoot, S. (2011). Weyl's scale invariance for standard model, renormalizability and zero cosmological constant. *Classical and Quantum Gravity*, 28, 145014.
- Noether, E. (1918). Invariante variationsprobleme. *Göttinger nachrichten* pp. 235–257. In *Gesamtelte abhandlungen* (vol. 1, pp. 770ff). Berlin: Springer.
- Novello, M. (1969). Dirac's equation in a Weyl space. *Nuovo Cimento A*, 94(4), 954–960.
- Novello, M., & Heintzmann, H. (1983). Weyl integrable space-time: A model for the cosmos? *Physics Letters A*, 98(1), 10–11.
- Novello, M., Oliveira, L. A. R., Salim, J. M., & Elbaz, E. (1992). Geometrized instantons and the creation of the universe. *International Journal of Modern Physics D*, 1, 641–677.
- Obukhov, Y. (1982). Conformal invariance and space-time torsion. *Physics Letters A*, 90, 13–16.
- Ohanian, H. (2016). Weyl gauge-vector and complex dilaton scalar for conformal symmetry and its breaking. *General Relativity and Gravity*, 48(25). <https://doi.org/10.1007/s10714-016-2023-8>. arXiv:1502.00020.
- Omote, M. (1971). Scale transformations of the second kind and the Weyl space-time. *Lettere al Nuovo Cimento*, 2(2), 58–60.
- Omote, M. (1974). Remarks on the local-scale-invariant gravitational theory. *Lettere al Nuovo Cimento*, 10(2), 33–37.
- O'RaiFeartaigh, L. (1997). *The dawning of gauge theory*. Princeton: Princeton University Press.
- Ornea, L. (2001). Weyl structures in quaternionic geometry. A state of the art. In E. Barletta (Ed.), *Selected topics in geometry and mathematical physics* (vol. 1, pp. 43–80). Potenza: University of degli Studi della Basilicata. arXiv:math/0105041.

- Padmanabhan, T. (1989). Quantum cosmology – the story so far. In B. R. Iyer, N. Mukunda, & C. V. Vishveshwara (Eds.), *Gravitation, gauge theories and the early universe* (pp. 373–404). Dordrecht: Kluwer
- Passon, O. (2004). *Bohmsche Mechanik. Eine Einführung in die deterministische Interpretation der Quantenmechanik*. Frankfurt/Main: Harri Deutsch.
- Passon, O. (2015). *Nicht-Kollaps-Interpretationen der Quantentheorie*. In C. Friebe, M. Kuhlmann, H. Lyre, P. M. Näger, O. Passon, & M. Stöckler (Eds.), *Philosophie der quantenphysik* (pp. 178–224). Berlin: Springer.
- Pauli, W. (1921). Relativitätstheorie. In *Encyklopädie der Mathematischen Wissenschaften* (vol. 5, pp. 539–775). Leipzig: Teubner. Collected Papers I, 1–237. Reprint edited and commented by D. Giulini, Berlin etc. Springer 2000.
- Pauli, W. (1940). Über die Invarianz der Dirac'schen Wellengleichungen gegenüber Ähnlichkeitstransformationen des Linienelementes im Fall verschwindender Ruhmasse. *Helvetica Physica Acta*, 13, 204–208. In (Pauli 1964, II, 918–922).
- Pauli, W. (1964). *Collected scientific papers*, R. Kronig, V. F. Weisskopf (Eds.). New York: Wiley.
- Pawłowski, M. (1990). Can gravity do what the Higgs does? Preprint IC/90/454.
- Pawłowski, M., & Rączka, R. (1994a). Mass generation in the standard model without dynamical Higgs field. Preprint. hep-th/9403303.
- Pawłowski, M., & Rączka, R. (1994b). A unified conformal model for fundamental interactions without dynamical Higgs field. *Foundations of Physics*, 24, 1305–1327. ILAS 4/94 hep-th/9407137.
- Pawłowski, M., & Rączka, R. (1995a). A Higgs-free model for fundamental interactions and its implications. Preprint. ILAS/EP-1-1995.
- Pawłowski, M., & Rączka, R. (1995b). A Higgs-free model for fundamental interactions. Part I: Formulation of the model. In J. Bertrand, M. Flato, J.-P. Gazeau, M. Irac-Astaud, & D. Sternheimer (Eds.), *Modern group theoretical methods in physics* (pp. 221–232). Springer Science+Business Media: Dordrecht. Preprint ILAS/EP-3-1995, hep-ph/9503269.
- Penrose, R. (1965). Zero rest-mass fields including gravitation: asymptotic behaviour. *Proceedings Royal Society London A*, 284, 159–203.
- Penrose, R. (2006). Before the big bang: An outrageous perspective and its implications for particle physics. In *Proceedings of EPAC 2006, Edinburgh, Scotland*. <http://accelconf.web.cern.ch/accelconf/e06/PAPERS/THESPA01.PDF>.
- Perlick, V. (1987). Characterization of standard clocks by means of light rays and freely falling particles'. *General Relativity and Gravitation*, 19, 1059–1073.
- Perlick, V. (1989). *Zur Kinematik Weylscher Raum-Zeit-Modelle*. Dissertationsschrift TU Berlin.
- Perlick, V. (1991). Observer fields in Weylian spacetime models. *Classical and Quantum Gravity*, 8, 1369–1385.
- Pfister, H. (2004). Newton's first law revisited. *Foundations of Physics Letters*, 17, 49–64.
- Pickering, A. (1988). *Constructing quarks*. Edinburgh: Edinburgh University Press.
- Pucheu, M. L., Almeida, T. S., & Romero, C. (2014). A geometrical approach to Brans-Dicke theory. In C. M. Gonzales, J. E. M. Aguiar, & L. M. R. Barrera (Eds.), *Accelerated Cosmic Expansion. Astrophysics and space science proceedings* (vol. 38, pp. 33–41). Berlin: Springer
- Pucheu, M. L., Alves, F. A. P., Barreto, A. B., & Romero, C. (2016). Cosmological models in Weyl geometric scalar tensor theory. *Physical Review D*, 6, 064010. arXiv:1602.06966.
- Quiros, I. (2000a). Dual geometries and spacetime singularities. *Physical Review D*, 61, 124026.
- Quiros, I. (2000b). Transformations of units and world's geometry. Preprint. gr-qc/0004014.
- Quiros, I. (2013). Scale invariance and broken electroweak symmetry may coexist together. Preprint. arXiv:1312.1018.
- Quiros, I. (2014a). Scale invariance: fake appearances. Preprint. arXiv:1405.6668.
- Quiros, I. (2014b). Scale invariant theory of gravity and the standard model of particles. Preprint. arXiv:1401.2643.
- Quiros, I., Bonal, R., & Cardenas, R. (2000). Brans-Dicke-type theories and avoidance of the cosmological singularity. *Physical Review D*, 62, 044042.

- Quiros, I., García-Salcedo, R., Madriz Aguilar, J., & Matos, T. (2013). The conformal transformations' controversy: what are we missing. *General Relativity and Gravitation*, 45, 489–518. arXiv:1108.5857.
- Ray, J. (1972). Lagrangian density for perfect fluids in general relativity. *Journal of Mathematical Physics*, 13(10), 1451–1453.
- Rievers, B., & Lämmerzahl, C. (2011). High precision thermal modeling of complex systems with application to the flyby and Pioneer anomaly. *Annalen der Physik*, 532(6), 439. arXiv:1104.3985.
- Rindler, W. (2006). *Relativity. Special, General, and Cosmological*. Oxford: Oxford University Press. 2nd ed. 2007.
- Romero, C., Fonseca-Neto, J.B., & Pucheu, M. L. (2011). General relativity and Weyl frames. *International Journal of Modern Physics A*, 26(22), 3721–3729. arXiv:1106.5543.
- Romero, C., Fonseca-Neto, J. B., & Pucheu, M. L. (2012). General relativity and Weyl geometry. *Classical and Quantum Gravity*, 29 (15), 155015. arXiv:1201.1469.
- Rosen, N. (1982). Weyl's geometry and physics. *Foundations of Physics*, 12, 213–248.
- Ruegg, H., & Ruiz-Altaba, M. (2003). The Stueckelberg field. Preprint. arXiv:hep-th/0304245.
- Ryckman, T. (2005). *The reign of relativity. Philosophy in physics 1915–1925*. Oxford: Oxford University Press.
- Salim, J. M., & Sautú, S. L. (1996). Gravitational theory in Weyl integrable spacetime. *Classical and Quantum Gravity*, 13 (2), 363–360.
- Sanders, R. H. (2010). *The dark matter problem. A historical perspective*. Cambridge: Cambridge University Press.
- Santamato, E. (1984a). Geometric derivation of the Schrödinger equation from classical mechanics in curved Weyl spaces. *Physical Review D*, 29, 216–222.
- Santamato, E. (1984b). Statistical interpretation of the Klein-Gordon equation in terms of the spacetime Weyl curvature. *Journal of Mathematical Physics* 25(8), 2477–2480.
- Santamato, E. (1985). Gauge-invariant statistical mechanics and average action principle for the Klein-Gordon particle in geometric quantum mechanics. *Physical Review D*, 32(10), 2615–2621.
- Santamato, E., & De Martini, F. (2013). Derivation of the Dirac equation by conformal differential geometry. *Foundations of Physics*, 43(5), 631–641. arxiv:1107.3168.
- Schneider, M. (2011). *Zwischen zwei Disziplinen. B.L. van der Waerden und die Entwicklung der Quantenmechanik*. Berlin: Springer.
- Scholz, E. (1999). Weyl and the theory of connections. In *Gray (1999)*. pp. 260–284.
- Scholz, E. (Ed.) (2001). *Hermann Weyl's Raum - Zeit - Materie and a general introduction to his scientific work*. Basel: Birkhäuser.
- Scholz, E. (2005a). Einstein-Weyl models of cosmology. In J. Renn (Ed.), *Albert Einstein. 100 authors for Einstein* (pp. 394–397). Weinheim: Wiley-VCH.
- Scholz, E. (2005b). On the geometry of cosmological model building. Preprint. arXiv:gr-qc/0511113.
- Scholz, E. (2009). Cosmological spacetimes balanced by a Weyl geometric scale covariant scalar field. *Foundations of Physics*, 39, 45–72. arXiv:0805.2557.
- Scholz, E. (2011a). Weyl geometric gravity and electroweak symmetry 'breaking'. *Annalen der Physik*, 523, 507–530. arxiv.org/abs/1102.3478.
- Scholz, E. (2011b). Weyl's scale gauge geometry in late 20th century physics. Preprint. arXiv:1111.3220.
- Scholz, E. (2016a). MOND-like acceleration in integrable Weyl geometric gravity. *Foundations of Physics*, 46, 176–208. arXiv:1412.0430.
- Scholz, E. (2016b). Clusters of galaxies in a Weyl geometric approach to gravity. *Journal of Gravity*, 46, 9706704. arXiv:1506.09138.
- Scholz, E. (2017). Paving the way for transitions – a case for Weyl geometry. In (Lehmkuhl et al. 2017, pp. 171–224). arXiv:1206.1559.

- Schouten, J. A. (1924). *Der Ricci-Kalkül. Eine Einführung in die neueren Methoden und problem der mehrdimensionalen differentialgeometrie. Die grundlehren der mathematischen wissenshaften* (vol. 10). Berlin: Springer.
- Schouten, J. A. (1954). *Ricci calculus* (2nd ed.). Berlin: Springer.
- Sciama, D. W. (1962). On the analogy between charge and spin in general relativity. In *Recent developments in general relativity Festschrift for L. Infeld* (pp. 415–439). Oxford and Warsaw: Pergamon and PWN. In (Blagojević/Hehl 2013, chap. 4).
- Shaposhnikov, M., & Zenhäusern, D. (2009a). Quantum scale invariance, cosmological constant and hierarchy problem. *Physics Letters B*, 671, 162–166. arXiv:0809.3406.
- Shaposhnikov, M., & Zenhäusern, D. (2009b). “Scale invariance, unimodular gravity and dark energy. *Physics Letters B*, 671, 187–192. arXiv:0809.3395.
- Sharpe, R. W. (1997). *Differential geometry: Cartan’s generalization of Klein’s Erlangen program*. Berlin: Springer.
- Shojai, A. (2000). Quantum gravity and cosmology. *International Journal of Modern Physics A*, 15(2), 1757–1771.
- Shojai, F., & Golshani, M. (1998). On the geometrization of Bohmian quantum mechanics: A new approach to quantum gravity. *International Journal of Modern Physics A*, 13(4), 677–693.
- Shojai, A., & Shojai, F. (2000). Nonminimal scalar-tensor theories and quantum gravity. *International Journal of Modern Physics A*, 15(13), 1859–1868.
- Shojai, F., & Shojai, A. (2001). About some problems raised by the relativistic form of de-Broglie-Bohm theory of pilot wave. *Physica Scripta*, 54(5), 413–416. arXiv:gr-qc/0404102.
- Shojai, F., & Shojai, A. (2003). On the relation of Weyl geometry and Bohmian quantum mechanics. *Gravitation and Cosmology*, 9(3), 163ff. Max Planck Institute for Gravitational Physics, Preprint AEI-2002-060. arXiv:gr-qc/0306099.
- Shojai, F., & Shojai, A. (2004). Understanding quantum theory in terms of geometry. Preprint. arXiv:gr-qc/0404102.
- Shojai, F., Shojai, A., & Golshani, M. (1998a). Conformal transformations and quantum gravity. *Modern Physics Letters A*, 13 (34), 2725–2729.
- Shojai, F., Shojai, A., & Golshani, M. (1998b). Scalar tensor theories and quantum gravity. *Modern Physics Letters A*, 13(36), 1915–2922.
- Shojai, A., Shojai, F., & Golshani, M. (1998c). Nonlocal effects in quantum gravity. *Modern Physics Letters A*, 13(37), 2965–2969.
- Simms, D. (1978). On the Schrödinger equation given by geometric quantization. In K. Bleuler, H. R. Petry, & A. Reetz (Eds.), *Differential geometrical methods in mathematical physics II* (vol. 676, pp. 351–356) Lecture notes in mathematics. Berlin: Springer.
- Smolin, L. (1979). Towards a theory of spacetime structure at very short distances. *Nuclear Physics B*, 160, 253–268.
- Śniatycki, J. (1980). *Geometric quantization and quantum mechanics*. Berlin: Springer.
- Souriau, J.-M. (1966). Quantification géométrique. *Communications in Mathematical Physics*, 1, 374–398.
- Souriau, J.-M. (1970). *Structure des systèmes dynamiques*. Paris: Duno. English as (Souriau 1997).
- Souriau, J.-M. (1997). *Structure of dynamical systems. A symplectic view of physics*. Berlin: Springer. Translated from (Souriau 1970) by C.-H. Cushman-de Vpasseonries.
- Steinhardt, P., & Turok, N. (2002). Cosmic evolution in a cyclic universe. *Physical Review D*, 65(12), 126003. arXiv:hep-th/0111098.
- Stoeltzner, M. (2014). Higgs models and other stories about mass generation. *Journal for the General Philosophy of Science*, 45, 369–386.
- Tann, H. (1998). *Einbettung der quantentheorie eines skalarfeldes in eine Weyl geometrie — Weyl symmetrie und ihre brechung*. München: Utz.
- Tonnellat, M.-A. (1965). *Les Théories unitaires de l’électromagnétisme et de la gravitation*. Paris: Gauthier-Villars.
- Trautman, A. (1972). On the Einstein-Cartan equations I, II. *Bulletin Academie Polonaise des Sciences, Série des sciences math., astr. et phys.*, 20, 185–191, 503.

- Trautman, A. (1973). On the structure of the Einstein-Cartan equations. *Symposia Mathematica*, 12, 139–162. Relatività convegno del Febbraio del 1972.
- Trautman, A. (2006). Einstein-Cartan theory. In J.-P. Francoise, G. L. Naber, & S. T. Tsou (Eds.), *Encyclopedia of mathematical physics* (vol. 2, pp. 189–195). Oxford: Elsevier. In (Blagojević/Hehl 2013, chap. 4).
- Trautman, A. (2012). Editorial note to J. Ehlers, F. A. E. Pirani and A. Schild, The geometry of free fall and light propagation. *General Relativity and Gravity*, 44(1), 1581–1586.
- Utiyama, R. (1956). Invariant theoretical interpretation of interaction. *Physical Review*, 101(5), 1597–1607.
- Utiyama, R. (1973). On Weyl's gauge field. *Progress of Theoretical Physics*, 50, 2028–2090.
- Utiyama, R. (1975a). On Weyl's gauge field. *General Relativity and Gravitation*, 6, 41–47.
- Utiyama, R. (1975b). On Weyl's gauge field II. *Progress of Theoretical Physics*, 53, 565–574.
- Vizgin, V. (1994). *Unified Field Theories in the First Third of the 20th Century*. Basel: Birkhäuser. Translated from the Russian by J. B. Barbour.
- Weinberg, S. (1972). *Gravitation and cosmology*. New York: Wiley.
- Weyl, H. (1918a). Gravitation und Elektrizität. *Sitzungsberichte der königlich preußischen akademie der wissenschaften zu Berlin* (pp. 465–480). In (Weyl 1968, II, 29–42), English in (O'Raifeartaigh 1997, 24–37).
- Weyl, H. (1918b). *Raum, - Zeit - Materie. vorlesungen über allgemeine relativitätstheorie*. Berlin: Springer. Other editions: ²1919, ³1919, ⁴1921, ⁵1923, ⁶1970, ⁷1988, ⁸1993. English and French translations from the 4th ed. in 1922.
- Weyl, H. (1918c). Reine Infinitesimalgeometrie. *Mathematische Zeitschrift*, 2, 384–411. In (Weyl 1968, II, 1–28).
- Weyl, H. (1920). Letter H. Weyl to F. Klein, December 28, 1920. *Nachlass F. Klein Universitätsbibliothek Göttingen Codex Ms Klein 12*, 297.
- Weyl, H. (1921). Zur Infinitesimalgeometrie: Einordnung der projektiven und der konformen Auffassung. *Nachrichten Göttinger gesellschaft der wissenschaften* (pp. 99–112). In (Weyl 1968, II, 195–207).
- Weyl, H. (1922). *Space – Time – Matter*. Translated from the 4th German edition by H. Brose. London: Methuen. Reprint New York: Dover 1952.
- Weyl, H. (1949). *Philosophy of Mathematics and Natural Science*. Princeton: Princeton University Press. ²1950, ³2009.
- Weyl, H. (1949/2016). Similarity and congruence: a chapter in the epistemology of science. ETH Bibliothek, Hs 91a:31. Published in (Weyl 1955, 3rd edition, 153–166).
- Weyl, H. (1955). *Symmetrie*. Ins Deutsche übersetzt von Lulu Bechtolsheim. Basel/Berlin: Birkhäuser/Springer. ²1981, 3rd edition *Ergänzt durch einen Text aus dem Nachlass 'Symmetry and congruence'*, ed. D. Giulini et al. 2016: Springer.
- Weyl, H. (1968). *Gesammelte Abhandlungen*. K. Chandrasekharan (Ed.), vol. 4. Berlin: Springer.
- Wood, W. R., & Papini, G. (1992). Breaking Weyl invariance in the interior of a bubble. *Physical Review D*, 45, 3617–3627.
- Wood, W. R., & Papini, G. (1997). A geometric approach to the quantum mechanics of de Broglie and Vigier. In S. Jeffers, S. Roy, J.-P. Vigiér, & G. Hunter (Eds.), *The Present Status of the Quantum Theory of Light. Proceedings in the Honour of Jean-Pierre Vigiér* (pp. 247–258). Dordrecht: Kluwer. arXiv:gr-qc/9612042.
- Woodhouse, N. M. J. (1991). *Geometric quantization*. Oxford: Clarendon.
- Wu, C.-L. (2004). Conformal scaling gauge symmetry and inflationary universe. *International Journal of Modern Physics A*, 20, 811ff. arXiv:astro-ph/0607064.
- Yang, C.-N. (1980). Einstein's impact on theoretical physics. *Physics Today*, 33(6), 42–49. In (Yang 1983, 563–567).
- Yang, C. N. (1983). *Selected papers 1945–1980. With commentary*. San Francisco: Freeman.
- Yang, C. N., & Mills, R. (1954). Conservation of isotopic spin and isotopic gauge invariance. *Physical Review*, 96, 191–195. In (Yang 1983, 172–176)
- Yuan, F.-F., & Huang, Y.-C. (2013). A modified variational principle for gravity in Weyl geometry. *Classical and Quantum Gravity*, 30(19), 195008. arXiv:1301.1316.

- Zee, A. (1979). Broken-symmetric theory of gravity. *Physical Review Letters*, *42*, 417–421.
- Zee, A. (1982). A theory of gravity based on the Weyl-Eddington action. *Physics Letters B*, *109*, 183–186.
- Zee, A. (1983). Einstein gravity emerging from quantum Weyl gravity. *Annals of Physics*, *151*, 431–443.

Part IV
Mathematical Motifs in General Relativity
and Beyond

Chapter 12

Matter from Space



Domenico Giulini

12.1 Introduction

Towards the end of his famous habilitation address, delivered on June 10, 1854, to the Philosophical Faculty of the University of Göttingen, Bernhard Riemann applied his mathematical ideas to physical space and developed the idea that it, even though of Euclidean appearance at macroscopic scales, may well have a non-Euclidean geometric structure in the sense of variable curvature if resolved below some yet unspecified microscopic scale. It is remarkable that in this connection he stressed that the measure for geometric ratios ('Massverhältnisse') would already be encoded in the very notion of space itself if the latter were considered to be a discrete entity, whereas in the continuous case, the geometry must be regarded as being a contingent structure that depends on 'acting forces'.¹

This suggestion was seized and radicalised by William Kingdon Clifford (Figure 12.1), who in his paper 'On the Space-Theory of Matter', read to the Cambridge Philosophical Society on February 21, 1870, took up the tough stance that *all* material properties and happenings may eventually be explained in terms of the curvature of space and its changes. In this seminal paper, the 24-year-old Clifford said (Clifford 1982, reprinted in Pesic 2007, 71):

¹'Es muß also entweder das dem Raume zugrunde liegende Wirkliche eine diskrete Mannigfaltigkeit bilden, oder der Grund der Maßverhältnisse außerhalb, in darauf wirkenden bindenden Kräften gesucht werden.' (Riemann 1869/1919, 20)

D. Giulini (✉)

Institute for Theoretical Physics, Leibniz University Hannover, Appelstrasse 2, 30167 Hannover, Germany

e-mail: giulini@itp.uni-hannover.de

Fig. 12.1 Replica (by John Collier) of the portrait of William Kingdon Clifford at the London National Portrait Gallery



'I wish here to indicate a manner in which these speculations [Riemann's] may be applied to the investigation of physical phenomena. I hold in fact:

1. That small portions of space *are* in fact of a nature analogous to little hills on a surface which is on the average flat; namely, that the ordinary laws of geometry are not valid in them.
2. That this property of being curved or distorted is continually being passed from one portion of space to another after the manner of a wave.
3. That this variation of the curvature of space is what really happens in that phenomenon which we call the *motion of matter*, whether ponderable or ethereal.
4. That in the physical world nothing else takes place but this variation, subject (possibly) to the law of continuity.'

In this contribution I wish to explain and comment on the status of this programme within general relativity. This is not to suggest that present-day physics offers even the slightest hope that this programme – understood in its radical sense – could succeed. But certain aspects of it certainly are realised, sometimes in a rather surprising fashion, and this is what I wish to talk about here.

That matter-free physical space should have physical properties at all seems to be quite against the view of Leibniz, Mach, and their modern followers, according to which space is a relational concept whose ontological status derives from that of the fundamental constituents of matter whose relations are considered. But at the same time it also seems to be a straightforward consequence of modern field theory, according to which fundamental fields are directly associated with space (or spacetime) rather than any space-filling material substance. Once the latter view is adopted, there seems to be no good reason to neglect the field that describes the geometry of space. This situation was frequently and eloquently described by Einstein, who empathetically wrote about the difficulties that one encounters in attempting to mentally emancipate the notion of a field from the idea of a substantial carrier whose physical states the field may describe. In doing this, the field describes



'People slowly accustomed themselves to the idea that the physical states of space itself were the final physical reality...' (A. Einstein, 1929)

Fig. 12.2 Cartoon of 1929 in *The New Yorker* by its first art editor Rea Irvin. Credit: The New Yorker © Condé Nast

the states of space itself, so that space becomes a dynamical agent, albeit one to which standard kinematical states of motion cannot be attributed, as Einstein stressed, e.g. in his 1920 Leiden address 'Äther und Relativitätstheorie' (Einstein 1920). A famous and amusing cartoon is shown in Figure 12.2, whose caption quotes Einstein expressing a view close to that of Clifford's.

12.2 Geometrodynamics

The field equations of general relativity with cosmological constant Λ read ($\kappa = 8\pi G/c^4$, where G is Newton's constant)

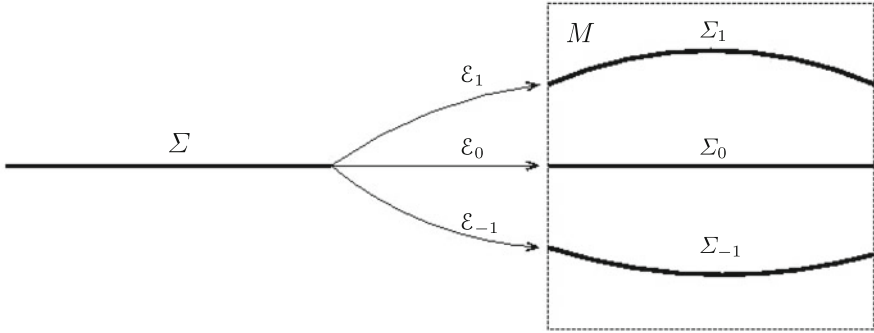


Fig. 12.3 Spacetime M is foliated by a one-parameter family of space-like embeddings of the 3-manifold Σ . Here the image Σ_1 of Σ under $\mathcal{E}_{t=1}$ lies to the future (above) and $\Sigma_{-1} := \mathcal{E}_{t=-1}$ to the past (below) of $\Sigma_0 := \mathcal{E}_{t=0}(\Sigma)$

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R + g_{\mu\nu}\Lambda = \kappa T_{\mu\nu} . \tag{12.1}$$

They form a system of ten quasilinear partial differential equations for the ten components $g_{\mu\nu}$ of the spacetime metric. These equations may be cast into the form of evolution equations. More precisely, the system (12.1) may be decomposed into a subsystem of four underdetermined elliptic equations that merely constrain the initial data (the so-called constraints) and a complementary subsystem of six underdetermined hyperbolic equations that drives the evolution. (The underdetermination is in both cases a consequence of diffeomorphism invariance.) This split is made possible by foliating spacetime M into three-dimensional space-like leaves Σ_t via a one-parameter family of embeddings $\mathcal{E}_t : \Sigma \hookrightarrow M$ with images $\mathcal{E}_t(\Sigma) = \Sigma_t \subset M$ (see Figure 12.3). The object that undergoes evolution in this picture is the three-dimensional Riemannian manifold (Σ, h) whose metric at time t is $h_t = \mathcal{E}_t^*g$, where g is the spacetime metric. In this evolutionary picture, spacetime appears as space’s history.

12.2.1 Hypersurface Kinematics

Let us be more precise on what it means to say that spacetime is considered as the trajectory (history) of space. Let $\text{Emb}(\Sigma, M)$ denote the space of smooth space-like embeddings $\Sigma \rightarrow M$. We consider a curve $\mathbb{R} \ni t \rightarrow \mathcal{E}_t \in \text{Emb}(\Sigma, M)$ corresponding to a one-parameter family of smooth embeddings with space-like images. We assume the images $\mathcal{E}_t(\Sigma) =: \Sigma_t \subset M$ to be mutually disjoint and moreover $\hat{\mathcal{E}} : \mathbb{R} \times \Sigma \rightarrow M, (t, p) \mapsto \mathcal{E}_t(p)$, to be an embedding. (It is sometimes found convenient to relax this condition, but this is of no importance here). The Lorentz manifold $(\mathbb{R} \times \Sigma, \mathcal{E}^*g)$ may now be taken as (\mathcal{E} -dependent) representative of M (or at least some open part of it) on which the leaves of the

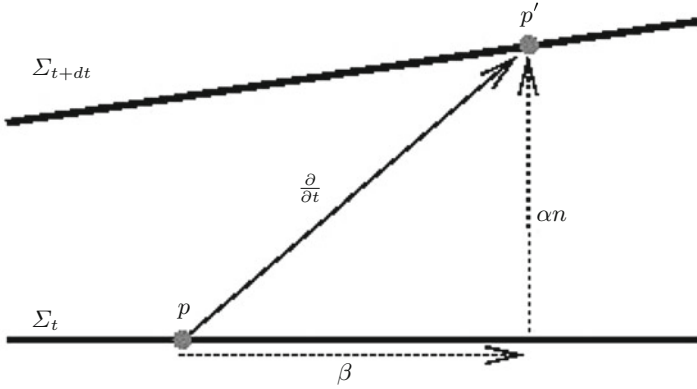


Fig. 12.4 For $q \in \Sigma$, the image points $p = \mathcal{E}_t(q)$ and $p' = \mathcal{E}_{t+dt}(q)$ are connected by the vector $\partial/\partial t|_p$ whose components tangential and normal to Σ_t are β (three functions) and αn (one function), respectively

above foliation simply correspond to the $t = \text{const.}$ hypersurfaces. Let n denote a field of normalised time-like vectors normal to these leaves. n is unique up to orientation, so that the choice of n amounts to picking a ‘future direction’.

The tangent vector $d\mathcal{E}_t/dt|_{t=0}$ at $\mathcal{E}_0 \in \text{Emb}(\Sigma, M)$ corresponds to a vector field over \mathcal{E}_0 (i.e. section in $T(M)|_{\mathcal{E}_0(\Sigma)}$), given by

$$\frac{d\mathcal{E}_t(p)}{dt}\Big|_{t=0} =: \frac{\partial}{\partial t}\Big|_{\mathcal{E}_0(p)} = \alpha n + \beta \tag{12.2}$$

with components (α, β) normal and tangential to $\Sigma_0 \subset M$. The functions α (one function), usually called the *lapse function*, and β (3 functions), usually called the *shift vector field*, combine the four-function worth of arbitrariness in moving the hypersurface Σ in spacetime (see Figure 12.4).

Conversely, each vector field V on M defines a vector field $X(V)$ on $\text{Emb}(\Sigma, M)$, corresponding to the left action of $\text{Diff}(M)$ on $\text{Emb}(\Sigma, M)$ given by composition. In local coordinates y^μ on M and x^k on Σ , it can be written as

$$X(V) = \int_{\Sigma} d^3x V^\mu(y(x)) \frac{\delta}{\delta y^\mu(x)}. \tag{12.3}$$

One easily verifies that $X : V \mapsto X(V)$ is a Lie homomorphism:

$$[X(V), X(W)] = X([V, W]). \tag{12.4}$$

Alternatively, decomposing (12.3) into normal and tangential components with respect to the leaves of the embedding at which the tangent-vector field to $\text{Emb}(\Sigma, M)$ is evaluated yields an embedding-dependent parametrisation of $X(V)$ in terms of (α, β) :

$$X(\alpha, \beta) = \int_{\Sigma} d^3x \left(\alpha(x)n^\mu[y](x) + \beta^m(x)\partial_m y^\mu(x) \right) \frac{\delta}{\delta y^\mu(x)}, \tag{12.5}$$

where y in square brackets indicates the functional dependence of n on the embedding. The functional derivatives of n with respect to y can be computed (see the Appendix of Teitelboim (1973)) from which the commutator of deformation generators follows:

$$[X(\alpha_1, \beta_1), X(\alpha_2, \beta_2)] = -X(\alpha', \beta'), \tag{12.6}$$

where

$$\alpha' = \beta_1(\alpha_2) - \beta_2(\alpha_1), \tag{12.7a}$$

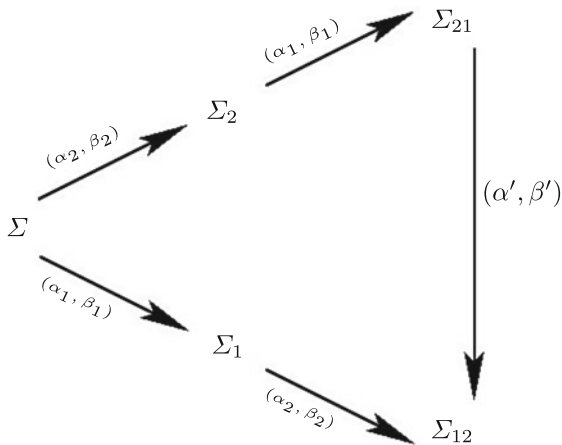
$$\beta' = [\beta_1, \beta_2] + \sigma\alpha_1 \text{grad}_h(\alpha_2) - \sigma\alpha_2 \text{grad}_h(\alpha_1). \tag{12.7b}$$

Here we left open whether spacetime M is Lorentzian ($\sigma = 1$) or Euclidean ($\sigma = -1$), in order to keep track of how the signature of spacetime ($-\sigma, +, +, +$) enters. Note that the h -dependent gradient field for the scalar function α is given by $\text{grad}_h(\alpha) = (h^{ab}\partial_b\alpha)\partial_a$. The geometric idea behind (12.7) is summarised in Figure 12.5.

12.2.2 Hamiltonian Geometrodynamics

The idea of Hamiltonian geometrodynamics is to realise these relations in terms of a Hamiltonian system on the phase space of physical fields. The most simple case is that where the latter merely include the spatial metric h on Σ , so that the

Fig. 12.5 An (infinitesimal) hypersurface deformation with parameters (α_1, β_1) that maps $\Sigma \mapsto \Sigma_1$, followed by one with parameters (α_2, β_2) that maps $\Sigma_1 \mapsto \Sigma_{12}$ differs by one with parameters (α', β') given by (12.7) from that in which the maps with the same parameters are composed in the opposite order



phase space is the cotangent bundle $T^*\text{Riem}(\Sigma)$ over $\text{Riem}(\Sigma)$. One then seeks a correspondence that associates to each pair (α, β) of lapse and shift a real-valued function on phase space:

$$(\alpha, \beta) \mapsto (H(\alpha, \beta) : T^*\text{Riem}(\Sigma) \rightarrow \mathbb{R}), \quad (12.8)$$

where

$$H(\alpha, \beta)[h, \pi] := \int_{\Sigma} d^3x (\alpha(x) \mathcal{H}[h, \pi](x) + h_{ab}(x) \beta^a(x) \mathcal{D}^b[h, \pi](x)), \quad (12.9)$$

with integrands $\mathcal{H}[h, \pi](x)$ and $\mathcal{D}^b[h, \pi](x)$ yet to be determined. H should be regarded as distribution (here the test functions are α and β^a) with values in real-valued functions on $T^*\text{Riem}(\Sigma)$. Now, the essential requirement is that the Poisson brackets between the $H(\alpha, \beta)$ are, up to a minus sign,² as in (12.6)–(12.7):

$$\{H(\alpha_1, \beta_1), H(\alpha_2, \beta_2)\} = H(\alpha', \beta'). \quad (12.10)$$

For the integration of canonical initial data (h, π) with Hamiltonian $H(\alpha, \beta)$, we need to specify by hand a one-parameter (representing parameter time t) family of lapse functions $t \mapsto \alpha_t(x) = \alpha(t, x)$ and shift vector fields $t \mapsto \beta_t(x) = \beta(t, x)$. It is now clear that this freedom just corresponds to the freedom to foliate the spacetime to be constructed. The Hamiltonian equations of motion contain only the unknown functions $h_t(x) = h(t, x)$ and $\pi_t(x) = \pi(t, x)$ and should be regarded as evolution equations (in terms of parameter time t) for the one-parameter families of tensor fields $t \mapsto h_t$ and $t \mapsto \pi_t$. Once the integration is performed, the solution gives rise to solution of Einstein's equation: If β_t^b is the one-form field corresponding to the vector field β_t via h_t , i.e. $\beta_t^b := h_t(\beta_t, \cdot)$, then the Lorentzian metric that satisfies Einstein's equation on the manifold $I \times \Sigma$, where I is the interval on the real line in which the parameter t takes its values, is given by

$$g = -(\alpha_t^2 - h_t(\beta_t, \beta_t))dt \otimes dt + \beta_t^b \otimes dt + dt \otimes \beta_t^b + h_t. \quad (12.11)$$

However, this integration may not start from any arbitrary set of initial data (h, π) . The data themselves need to satisfy a system of (underdetermined elliptic) partial differential equations, the so-called constraints. The reason for their existence as well as their analytic form will be explained in the next subsections.

²Due to the standard convention that the Hamiltonian action is defined as a *left* action, whereas the Lie bracket on a group is defined by the commutator of left-invariant vector fields which generate *right* translations.

12.2.3 Why Constraints

From (12.10) alone follows a remarkable uniqueness result as regards the analytical structure of $H(\alpha, \beta)$ as functional of (h, π) . Before stating it with all its hypotheses, we show why the constraints $\mathcal{H}[h, \pi] = 0$ and $\mathcal{D}^b[h, \pi] = 0$ must be imposed.

Consider the set of smooth real-valued functions on phase space, $F : T^*\text{Riem}(\Sigma) \rightarrow \mathbb{R}$. They are acted upon by all $H(\alpha, \beta)$ via Poisson bracketing: $F \mapsto \{F, H(\alpha, \beta)\}$. This defines a map from (α, β) into the derivations of phase-space functions. We require this map to also respect the commutation relation (12.10), that is, we require

$$\{\{F, H(\alpha_1, \beta_1)\}, H(\alpha_2, \beta_2)\} - \{\{F, H(\alpha_2, \beta_2)\}, H(\alpha_1, \beta_1)\} = \{F, H\}(\alpha', \beta'). \quad (12.12)$$

The crucial and somewhat subtle point to be observed here is the following: Up to now the parameters (α_1, β_1) and (α_2, β_2) were considered as given functions of $x \in \Sigma$, independent of the fields h and π , i.e. independent of the point of phase space. However, from (12.7b) we see that $\beta'(x)$ does depend on $h(x)$. This dependence should not give rise to extra terms $\propto \{F, \alpha'\}$ in the Poisson bracket, for, otherwise, the extra terms would prevent the map $(\alpha, \beta) \mapsto \{-, H(\alpha, \beta)\}$ from being a homomorphism from the algebraic structure of hypersurface deformations into the derivations of phase-space functions. This is necessary in order to interpret $\{-, H(\alpha, \beta)\}$ as a generator (on phase-space functions) of a *spacetime* evolution corresponding to a normal lapse α and tangential shift β . In other words, the evolution of observables from an initial hypersurface Σ_i to a final hypersurface Σ_f must be independent of the intermediate foliation ('integrability' or 'path independence' (Teitelboim 1973; Hojman et al. 1973, 1976)). Therefore we placed the parameters (α', β') outside the Poisson bracket on the right-hand side of (12.12), to indicate that no differentiation with respect to h, π should act on them.

To see that this requirement implies the constraints, rewrite the left-hand side of (12.12) in the form

$$\begin{aligned} & \{\{F, H(\alpha_1, \beta_1)\}, H(\alpha_2, \beta_2)\} - \{\{F, H(\alpha_2, \beta_2)\}, H(\alpha_1, \beta_1)\} \\ &= \{F, \{H(\alpha_1, \beta_1), H(\alpha_2, \beta_2)\}\} \\ &= \{F, H(\alpha', \beta')\} \\ &= \{F, H\}(\alpha', \beta') + H(\{F, \alpha'\}, \{F, \beta'\}), \end{aligned} \quad (12.13)$$

where the first equality follows from the Jacobi identity, the second from (12.10), and the third from the Leibniz rule. Hence the requirement (12.12) is equivalent to

$$H(\{F, \alpha'\}, \{F, \beta'\}) = 0 \quad (12.14)$$

for all phase-space functions F to be considered and all α', β' of the form (12.7). Since only β' depends on phase space, more precisely on h , this implies the

vanishing of the phase-space functions $H(0, \{F, \beta'\})$ for all F and all β' of the form (12.7b). This can be shown to imply $H(0, \beta) = 0$, i.e. $\mathcal{D}[h, \pi] = 0$. Now, in turn, for this to be preserved under all evolutions, we need $\{H(\alpha, \tilde{\beta}), H(0, \beta)\} = 0$, and hence in particular $\{H(\alpha, 0), H(0, \beta)\} = 0$ for all α, β , which implies $H(\alpha, 0) = 0$, i.e. $\mathcal{H}[h, \pi] = 0$. So we see that the constraints indeed follow from the required integrability condition.

Sometimes the constraints $H(\alpha, \beta) = 0$ are split into the *Hamiltonian (or scalar) constraints*, $H(\alpha, 0) = 0$, and the *diffeomorphisms (or vector) constraints*, $H(0, \beta) = 0$. The relations (12.10) with (12.7) then show that the vector constraints form a Lie-subalgebra which, because of $\{H(0, \beta), H(\alpha, 0)\} = H(\beta(\alpha), 0) \neq H(0, \beta')$, is not an ideal. This means that the Hamiltonian vector fields for the scalar constraints are not tangent to the surface of vanishing vector constraints, except where it intersects the surface of vanishing scalar constraints. This implies that the scalar constraints do not act on the solution space for the vector constraints, so that one simply cannot first reduce the vector constraints and then, on the solutions of that, search for solutions to the scalar constraints.

12.2.4 Uniqueness of Einstein's Geometrodynamics

It is sometimes stated that the relations (12.10) together with (12.7) determine the function $H(\alpha, \beta) : T^*\text{Riem}(\Sigma) \rightarrow \mathbb{R}$, i.e. the integrands $\mathcal{H}[h, \pi]$ and $\mathcal{D}[h, \pi]$, uniquely up to two free parameters, which may be identified with the gravitational and the cosmological constants. This is a mathematical overstatement if read literally, since the result can only be shown if certain additional assumptions are made concerning the action of $H(\alpha, \beta)$ on the basic variables h and π . The uniqueness result then obtained is still remarkable.

The first such assumption concerns the intended ('semantic' or 'physical') meaning of $H(0, \beta)$, namely, that the action of $H(0, \beta)$ on h or π is that of an infinitesimal spatial diffeomorphism of Σ . Hence it should be the spatial Lie derivative, L_β , applied to h or π . It then follows from the general Hamiltonian theory that $H(0, \beta)$ is given by the *momentum map* that maps the vector field β (viewed as element of the Lie algebra of the group of spatial diffeomorphisms) into the function on phase space given by the contraction of the momentum with the β -induced vector field $h \rightarrow L_\beta h$ on $\text{Riem}(\Sigma)$:

$$H(0, \beta) = \int_{\Sigma} d^3x \pi^{ab} (L_\beta h)_{ab} = -2 \int_{\Sigma} d^3x (\nabla_a \pi^{ab}) h_{bc} \beta^c. \quad (12.15)$$

Comparison with (12.9) yields

$$\mathcal{D}^b[h, \pi] = -2 \nabla_a \pi^{ab}. \quad (12.16)$$

The second assumption concerns the intended ('semantic' or 'physical') meaning of $H(\alpha, 0)$, namely, that $\{-, H(\alpha, 0)\}$ acting on h or π is that of an infinitesimal 'time-like' diffeomorphism of M normal to the leaves $\mathcal{E}_t(\Sigma)$. If M were given, it is easy to prove that we would have $L_{\alpha n}h = 2\alpha K$, where n is the time-like field of normals to the leaves $\mathcal{E}_t(\Sigma)$ and K is their extrinsic curvature. Hence one requires

$$\{h, H(\alpha, 0)\} = 2\alpha K. \quad (12.17)$$

Note that both sides are symmetric covariant tensor fields over Σ . The important fact to be observed here is that α appears without differentiation. This means that $H(\alpha, 0)$ is an ultralocal functional of π , which is further assumed to be a polynomial. Note that at this moment we do not assume any definite relation between π and K . Rather, this relation is a consequence of (12.17) once the analytic form of $H(\alpha, 0)$ is determined.

The Hamiltonian evolution so obtained is precisely that of general relativity (without matter) with two free parameters, which may be identified with the gravitational constant $\kappa = 8\pi G/c^4$ and the cosmological constant Λ . The proof of the theorem is given in Kuchar (1973), which improves on earlier versions (see Teitelboim 1973 and Hojman et al. 1976) in that the latter assume in addition that $\mathcal{H}[h, \pi]$ be an even function of π , corresponding to the requirement of time reversibility of the generated evolution. This was overcome in Kuchar (1973) by the clever move to write the condition set by $\{H(\alpha_1, 0), H(\alpha_2, 0)\} = H(0, \beta')$ (the right-hand side being already known) on $H(\alpha, 0)$ in terms of the corresponding Lagrangian functional L , which is then immediately seen to turn into a condition which is *linear* in L , so that terms with even powers in velocity decouple from those with odd powers. There is a slight topological subtlety remaining which is further discussed in Giulini (2009). The two points which are important for us here are:

1. *The dynamics of the gravitational field as given by Einstein's equations can be fully understood in term of the constraints.*
2. *Modulo some technical assumptions spelled out above, the constraints for pure gravity follow from the kinematical relation (12.10) with (12.7), once one specifies and gravitational phase space to be $T^*\text{Riem}(\Sigma)$, i.e. the gravitational configuration space to be $\text{Riem}(\Sigma)$.*

12.2.5 What the Constraints Look Like

Rather than writing down the constraints in terms of h and π , we shall use the simple relation between π and K that follows from (12.17) for given $H(\alpha, 0)$, the reason being that K has the simple interpretation as extrinsic curvature (also called second fundamental form) of the images of Σ in M , which is rather intuitive. From the determination of $H(\alpha, 0)$ the h -dependent relation between π and K , in terms of components, turns out to be

$$\pi^{ab} = \sqrt{\det\{h_{nm}\}} h^{ac} h^{bd} (K_{cd} - h_{cd} h^{ij} K_{ij}). \tag{12.18}$$

In terms of h and K , the constraints then assume the form

$$(h^{ac} h^{bd} K_{ab} K_{cd} - (h^{ab} K_{ab})^2) - (R(h) - 2\Lambda) = - (2\kappa)\rho, \tag{12.19a}$$

$$\nabla_b (K^{ab} - h^{ab} h^{nm} K_{nm}) = (c\kappa) j^a, \tag{12.19b}$$

The right-hand sides of both equations (12.19) are zero in the matter free (vacuum) case which we consider here. But we think it is instructive to know what it will be in the presence of matter: Here ρ and j represent the matter’s energy and momentum densities on Σ , respectively. Recall that the symmetry of the energy-momentum tensor for the matter implies that the momentum density is c^{-2} times the energy current density (energy per unit surface area and unit time). Moreover, $R(h)$ is the Ricci scalar for h , Λ is the cosmological constant, and $\kappa = 8\pi G/c^4$ as in (12.1).

The first bracket on the left-hand side of (12.19a) contains an h -dependent bilinear form in K . It can be seen as the kinetic term in the Hamiltonian of the gravitational field. Usually the kinetic term is positive definite, but this time it is not! Hence we wish to understand this bilinear form in more detail. In particular: Under what conditions on K is it positive or negative definite? This can be answered in terms of the eigenvalues of K . To make this precise, let \tilde{K} be the endomorphism field which is obtained from K by raising one index (which one does not matter due to symmetry) using h . We may now unambiguously speak of the eigenvalues of \tilde{K} , a triple for each space point. Each triple we collect in an eigenvalue vector $\lambda \in \mathbb{R}^3$. In terms of \tilde{K} , the bilinear form reads $\text{tr}(\tilde{K}^2) - (\text{tr}(\tilde{K}))^2$, which equals $\|\lambda\|^2 - (\lambda \cdot \mathbf{d})^2$ in terms of λ . Here the dot product and the norm are the usual ones in \mathbb{R}^3 and \mathbf{d} is the ‘diagonal vector’ with unit entries (1, 1, 1). Hence the bilinear form is positive definite iff³ the modulus of the cosine of the angle between λ and \mathbf{d} is less than $1/\sqrt{3}$ and negative definite iff it is greater than $1/\sqrt{3}$. In other words, the bilinear form is negative definite on those \tilde{K} whose eigenvalue vector lies in the interior of a double cone whose vertex is the origin, whose symmetry axis is the ‘diagonal’ generated by \mathbf{d} , and whose opening angle (angle between symmetry axis and boundary) is $\arccos(1/\sqrt{3}) \approx 54.7^\circ$. Note that this opening angle is just the one at which the boundary of each cone just contains all three positive or negative coordinate half-axes. It properly contains the maximal cones contained in the positive and negative octants, whose opening angle is $\arccos(\sqrt{2/3}) \approx 35.3^\circ$.⁴ Hence strictly positive or strictly negative eigenvalues of \tilde{K} imply a negative definite value of the bilinear form, but the converse is not true.⁵

³In this article, we use ‘iff’ as abbreviation for ‘if and only if’.

⁴The maximal cone touches the three 2-planes $\lambda_i = 0$ at the bisecting lines $\lambda_j = \lambda_k$, where i, j and k are any of the three cyclic permutations of 1, 2 and 3. Hence the cosine of the opening angle is the scalar product between $(1, 1, 1)\sqrt{3}$ and, say, $(1, 1, 0)\sqrt{2}$, which is $\sqrt{2/3}$.

⁵Had we done the very same analysis in terms of π rather than K , we would have found that in eigenvalue space (now of the endomorphism $\tilde{\pi}$) the opening angle of the cone inside which

The preceding discussion shows that the bilinear form is not of a definite nature. In fact, it is a $(1 + 5)$ – dimensional Lorentzian metric on the six-dimensional space of positive definite bilinear forms over a real three-dimensional vector space (the tangent space to Σ), which is known as the DeWitt metric since DeWitt’s seminal paper (DeWitt 1967) on canonical quantum gravity. Parametrising it by h_{ab} or (τ, r_{ab}) it can be written as

$$G^{abcd} dh_{ab} \otimes dh_{cd} = -(32/3) d\tau \otimes d\tau + \tau^2 \operatorname{tr}(r^{-1} dr \otimes r^{-1} dr), \quad (12.20a)$$

where

$$r_{ab} := [\det(h)]^{-1/3} h_{ab}, \quad \tau := [\det(h)]^{1/4}, \quad (12.20b)$$

and

$$G^{abcd} = \frac{1}{2} \sqrt{\det(h)} (h^{ac} h^{bd} + h^{ad} h^{bc} - 2h^{ab} h^{cd}). \quad (12.20c)$$

The form (12.20a) clearly reveals its geometric meaning as a warped-product metric of ‘cosmological type’ on the manifold $\mathbb{R} \times SL(3, \mathbb{R})/SO(3)$, where the five-dimensional homogeneous space $SL(3, \mathbb{R})/SO(3)$, parametrised by r_{ab} , carries its left invariant metric $\operatorname{tr}(r^{-1} dr \otimes r^{-1} dr) = r^{ac} r^{bd} dr_{ab} \otimes dr_{cd}$.

This pointwise Lorentzian metric induces a metric on the infinite-dimensional manifold $\operatorname{Riem}(\Sigma)$, known as Wheeler-DeWitt metric, through

$$\mathfrak{G}(k, \ell) = \int_{\Sigma} d^3x G^{abcd} k_{ab} \ell_{cd} \quad (12.21)$$

where the tensor fields k and ℓ are now considered as tangent vectors at $h \in \operatorname{Riem}(\Sigma)$. See Giulini (2009) and references therein for a recent review on geometric aspects associated with this metric and its role in geometrodynamics.

To end this brief sketch of geometrodynamics, let me just stress its (admittedly somewhat crude) analogy to relativistic point mechanics. The latter takes place in Minkowski space which is endowed with an absolute (i.e. non-dynamical) geometry through the Minkowski metric. Here the configuration space is $\operatorname{Riem}(\Sigma)$, which is also endowed with an absolute geometry through the Wheeler-DeWitt metric, although it is not true that the Einstein equations correspond to geodesic motion in it. However, the deviation from geodesic motion derives from a force that corresponds to a vector field on $\operatorname{Riem}(\Sigma)$ given by $-2(R_{ab} - \frac{1}{4} h_{ab} R)$, where R_{ab} and R are the Ricci tensor and scalar for h , respectively (Giulini 1995c).

the bilinear form is negative definite and outside which it is positive definite is now *precisely* the maximal one $\arccos(\sqrt{2/3})$ (see previous footnote). Indeed, rewriting the bilinear form in terms of π using (12.18), it is positively proportional to $\operatorname{tr}(\tilde{\pi}^2) - \frac{1}{2}(\operatorname{tr}(\tilde{\pi}))^2$. It is the one-half in front of the second term that causes this interesting coincidence.

12.2.6 Vacuum Data

Following Clifford's dictum, we shall in the following be interested in vacuum data, that is, data (h, K) that satisfy (12.19) for vanishing right-hand sides. Upon evolution these give rise to solutions $g_{\mu\nu}$ to Einstein's equations (12.1) for $T_{\mu\nu} = 0$.

An important nontrivial observation is that the system (12.19) does not impose any topological obstruction on Σ . That means that for any topological 3-manifold Σ , there are data (h, K) that satisfy (12.19) with vanishing right-hand side. This result can be understood as an immediate consequence of a famous theorem proved in Kazdan and Warner (1975), which states that *any* smooth function $f : \Sigma \rightarrow \mathbb{R}$ which is negative somewhere can be the scalar curvature for some Riemannian metric. Given that strong result, we may indeed always solve (12.19) for $\rho = 0$ and $j = 0$ as follows: First we make the Ansatz $K = \alpha h$ for some constant α and some $h \in \text{Riem}(\Sigma)$. This solves (12.19b), whatever α and h will be. Given the spacetime interpretation of K as extrinsic curvature, this means that the initial Σ will be a totally umbilic hypersurface in the spacetime M that is going to evolve from the data. Next we solve (12.19a) by fixing α so that $\alpha^2 > \Lambda/3$ and then choosing h so that $R(h) = 2\Lambda - 6\alpha^2$, which is possible by the result just cited because the right-hand side is negative by construction.

Simple but nevertheless very useful examples of vacuum data are provided by time-symmetric conformally flat ones. Time symmetry means that the initial 'velocity' of h vanishes, and hence that $K = 0$, so that (12.19b) is already satisfied. A vanishing extrinsic curvature is equivalent to saying that the hypersurface is totally geodesic, meaning that a geodesic in spacetime that initially starts on and tangent to $\Sigma \subset M$ will remain within Σ . This is to be expected since motions with vanishing initial velocity should be time-reflection symmetric, which here would imply the existence of an isometry of M (the history of space) that exchanges both sides of Σ in M and leaves Σ pointwise fixed. However, a fixed-point set of an isometry is always totally geodesic, for, if it were not, a geodesic starting on and tangent to Σ but taking off Σ eventually would be mapped by the isometry to a different geodesic with the same initial conditions, which contradicts the uniqueness theorem for solutions of the geodesic equation.

As time symmetry implies $K = 0$, we have automatically solved (12.19b) for $j = 0$. That h be conformally flat means that we may write $h = \phi^4 \delta$, where δ is the flat Euclidean metric on Σ and $\phi : \Sigma \rightarrow \mathbb{R}_+$ is a positive real-valued function. The remaining constraint (12.19a) for $\rho = 0$ then simply reduces to Laplace's equation for ϕ :

$$\Delta\phi = 0, \tag{12.22}$$

where Δ denotes the Laplace operator with respect to the flat metric δ .

Usually one seeks solutions so that (Σ, h) is a manifold with a finite number of asymptotically flat ends. One such end is then associated with 'spatial infinity', which really just means that the solution under consideration represents a quasi

isolated lump of geometry with a sufficiently large (compared to its own dimension) almost flat transition region to the ambient universe. According to Arnowitt, Deser and Misner (see their review Arnowitt et al. 1962), we can associate a (active gravitational) mass to each such end, which is defined as a limit of a flux integrals over 2-spheres pushed into the asymptotic region of the end in question. If the mass is measured in geometric units (i.e. it has the physical dimension of a length and is converted to a mass in ordinary units by multiplication with c^2/G), it is given by

$$m := \lim_{R \rightarrow \infty} \left\{ \frac{1}{16\pi} \int_{S_R} (\partial_j h_{ij} - \partial_i h_{jj}) n^i d\Omega \right\}, \tag{12.23}$$

where $S_R \subset \Sigma$ is a 2-sphere of radius R , outward-pointing normal n , and surface measure $d\Omega$. For later use, we note in passing that if the asymptotically flat spacetime is globally stationary (i.e. admits a time-like Killing field K), the overall mass can also be written in the following simple form, known as ‘Komar integral’ (Komar 1959):

$$m := \lim_{R \rightarrow \infty} \left\{ \frac{-1}{8\pi} \int_{S_R} \star dK^b \right\}, \tag{12.24}$$

where d denotes the exterior differential on spacetime, $K^b := g(K, \cdot)$ denotes the one form corresponding to K under the spacetime metric g (lowering the index) and \star is the Hodge-duality map. A similar expression exists for the overall angular momentum of a rotationally symmetric spacetime, as we shall see later.

The celebrated ‘positive-mass theorem’ states in the vacuum case that for any Riemannian metric h which satisfies the constraints (12.19) with $\rho = 0$ and $j = 0$ for some K has $m \geq 0$ for each asymptotically flat end.⁶ Moreover, $m = 0$ iff (Σ, h) is a spatial slice through Minkowski space. This already implies that the mass must be strictly positive if Σ is topologically different from \mathbb{R}^3 : Nontrivial topology implies nonzero positive mass! This is supported by the generalisation of the Penrose-Hawking singularity theorems due to Gannon (1975), which basically states that the geometric hypothesis of the existence of closed trapped surfaces in Σ in the former may be replaced by the purely topological hypothesis of Σ not being simply connected. This is our first example in GR of how attributes of matter (here mass) arise from pure geometry/topology.

⁶Note that the definition of the ADM mass (12.23) just depends on the Riemannian metric h and is independent of K . But for the theorem to hold, it is essential to require that h is such that there exists a K so that $(h$ and $K)$ satisfy the constraints. It is easy to write down metrics h with negative mass: Take, e.g. (12.29) with negative m for $r > r_* > m/2$, smoothly interpolated within $m/2 < r < r_*$ to, say, the flat metric in $r < m/2$. The positive mass theorem implies that for such a metric, no K can be found so that $(h$ and $K)$ satisfy the constraints.

12.2.7 Solution Strategies

A variety of methods exist to construct interesting solutions to (12.22). One of them is the ‘method of images’ known from electrostatics (Misner 1963). It is based on the conformal properties of the Laplace operator, which are as follows: Let $\Sigma = \mathbb{R}^3 - \{\mathbf{x}_0\}$ and δ its usual flat metric. Consider a sphere S_0 of radius r_0 centred at \mathbf{x}_0 . The ‘inversion at S_0 ’, denoted by $I_{(\mathbf{x}_0, r_0)}$, is a diffeomorphism of Σ that interchanges the exterior and the interior of $S_0 \subset \Sigma$ and leaves S_a pointwise fixed. In spherical polar coordinates centred at \mathbf{x}_0 , it takes the simple form

$$I_{(\mathbf{x}_0, r_0)}(r, \theta, \varphi) = (r_0^2/r, \theta, \varphi). \tag{12.25a}$$

There is a variant of this map that results from an additional antipodal reflection in the 2-spheres so that no fixed points exist:

$$I'_{(\mathbf{x}_0, r_0)}(r, \theta, \varphi) = (r_0^2/r, \pi - \theta, \varphi + \pi). \tag{12.25b}$$

Associated to each of these self-maps of Σ are self-maps $J_{(\mathbf{x}_0, r_0)}$ and $J'_{(\mathbf{x}_0, r_0)}$ of the set of smooth real-valued functions on Σ , given by

$$J_{(\mathbf{x}_0, r_0)}(f) = (r_0/r) (f \circ I_{(\mathbf{x}_0, r_0)}) \tag{12.26}$$

and likewise with $I'_{(\mathbf{x}_0, r_0)}$ exchanged for $I_{(\mathbf{x}_0, r_0)}$ on the right-hand side in case of $J'_{(\mathbf{x}_0, r_0)}$. Now, the point is that these maps obey the following simple composition laws with the Laplace operator (considered as self-map of the set of smooth functions on Σ):

$$\Delta \circ J_{(\mathbf{x}_0, r_0)} = (r_0/r)^4 J_{(\mathbf{x}_0, r_0)} \circ \Delta \tag{12.27}$$

and likewise with $J'_{(\mathbf{x}_0, r_0)}$ replacing $J_{(\mathbf{x}_0, r_0)}$. In particular, the last equation implies that $J_{(\mathbf{x}_0, r_0)}$ and $J'_{(\mathbf{x}_0, r_0)}$ map harmonic functions (i.e. functions ϕ satisfying $\Delta\phi = 0$) on Σ to harmonic functions on Σ . Note that Σ did not include the point \mathbf{x}_0 at which the sphere of inversion was centred. It is clear from (12.26) that the maps $J_{(\mathbf{x}_0, r_0)}$ will change the singular behaviour of the functions at \mathbf{x}_0 . For example, the image of the constant function $f \equiv 1$ under either $J_{(\mathbf{x}_0, r_0)}$ or $J'_{(\mathbf{x}_0, r_0)}$ is just the function $\mathbf{x} \mapsto r_0/\|\mathbf{x} - \mathbf{x}_0\|$, i.e. a pole of strength r_0 at \mathbf{x}_0 . Iterating once more, a pole of strength a located at \mathbf{a} is mapped via $J_{(\mathbf{x}_0, r_0)}$ resp. $J'_{(\mathbf{x}_0, r_0)}$ to a pole of strength $a/\|\mathbf{x}_0 - \mathbf{a}\|$ at $I_{(\mathbf{x}_0, r_0)}(\mathbf{a})$ resp. $I'_{(\mathbf{x}_0, r_0)}(\mathbf{a})$.

The general strategy is then as follows: Take a set $S_i, i = 1, \dots, N$, of N spheres with radii r_i and centres \mathbf{x}_i , so that each sphere S_i is disjoint from, and to the outside of, each other sphere $S_j, j \neq i$. Take the constant function $f \equiv 1$ and take the sum over the free group generated by all $J_{(\mathbf{x}_i, r_i)}$ (alternatively the $J'_{(\mathbf{x}_i, r_i)}$). This converges to an analytic function ϕ provided $(N - 1)r_*/d < 1$, where $r_* = \max\{r_1, \dots, r_N\}$ and d is the infimum of Euclidean distances from the centres \mathbf{x}_i to points on

the spheres S_j , $j \neq i$ (see Misner 1963 and Giulini 1990). By construction the function ϕ is then invariant under each inversion map $J_{(\mathbf{x}_i, r_i)}$ (alternatively $J'_{(\mathbf{x}_i, r_i)}$). Consequently, the maps $I_{(\mathbf{x}_i, r_i)}$ (alternatively $I'_{(\mathbf{x}_i, r_i)}$) are isometries of the metric $h := \phi^4 \delta$, which is defined on the manifold Σ which one obtains by removing from \mathbb{R}^3 the centres \mathbf{x}_i and all their image points under the free group generated by the inversions $I_{(\mathbf{x}_i, r_i)}$ or $I'_{(\mathbf{x}_i, r_i)}$. However, the topology of the manifold may be modified by suitable identifications using these isometries. For example, using $I'_{(\mathbf{x}_i, r_i)}$, we may excise the interiors of all spheres S_i and identify antipodal points on each remaining boundary component S_i . In this fashion, we obtain a manifold with one end, which is the connected sum of N real-projective spaces minus a point (the point at spatial infinity).

In general there are many topological options. Consider, for example, the simpler case of just two 2-spheres S_1 and S_2 of, say, equal radii, $r_1 = r_2$. We again excise their interiors and identify their boundaries. If we use the maps $I_{(\mathbf{x}_i, r_i)}$ for the data construction, we may identify S_1 with S_2 in an orientation reversing fashion (with respect to their induced orientations) so that the quotient space is orientable. This results in Misner’s wormhole (Misner 1959) whose data are often used in numerical studies of black-hole collisions. If instead we had used the maps $I'_{(\mathbf{x}_i, r_i)}$, we have two choices: either to identify antipodal points on each S_i separately, which results in the connected sum of two real-projective spaces, as explained above, or to identify S_1 with S_2 , but now in an orientation preserving fashion (with respect to their induced orientations) so that the resulting manifold is a non-orientable version of Misner’s wormhole discussed in Giulini (1990). The latter two manifolds are locally isometric but differ in their global topology, whereas they are not even locally isometric to the standard (orientable) Misner’s wormhole (Figure 12.6).

Let us turn to the simplest nontrivial example: a single black hole. It corresponds to the solution of (12.22) with a single pole at, say, $\mathbf{x}_0 = \mathbf{0}$ and asymptotic value $\phi \rightarrow 1$ for $r \rightarrow \infty$, where $r := \|\mathbf{x}\|$. Hence we have

$$\phi(\mathbf{x}) = 1 + \frac{m}{2r}. \tag{12.28}$$

It is easy to verify that the constant m just corresponds to the ADM mass defined via (12.23). The three-dimensional Riemannian manifold (Σ, h) is now given by $\Sigma = \mathbb{R}^3 - \{\mathbf{0}\}$ and the metric, in polar coordinates centred at the origin,

$$h = \left(1 + \frac{m}{2r}\right)^4 \underbrace{(dr^2 + r^2(d\theta^2 + \sin^2\theta d\varphi^2))}_{=\delta} \tag{12.29}$$

It allows for the two discrete isometries

$$I : (r, \theta, \varphi) \mapsto (m^2/4r, \theta, \varphi), \tag{12.30}$$

$$J : (r, \theta, \varphi) \mapsto (m^2/4r, \pi - \theta, \varphi + \pi). \tag{12.31}$$

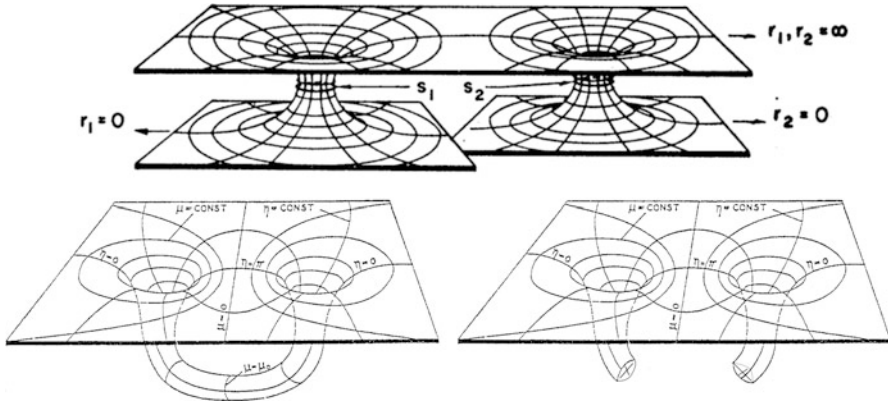


Fig. 12.6 Various topologies for data (Σ, h) representing two black holes momentarily at rest. The upper manifold has three asymptotically flat ends, one at spatial infinity and one each ‘inside’ the apparent horizons (= minimal surfaces) S_1, S_2 . The lower two manifolds have only one end each. The lower left manifold (wormhole) is topologically $S^1 \times S^2 - \{\text{point}\}$ the lower right $\mathbb{R}P^3 \# \mathbb{R}P^3 - \{\text{point}\}$, where $\#$ denotes connected sum. The crosswise arrows in the lower right picture indicate that the shown 2-sphere boundaries are closed off by antipodal identifications. The coordinates μ, η correspond to bispherical polar coordinates. No two of these three manifolds are locally isometric

The set of fixed points for I is the sphere $r = m/2$, whereas J acts freely (without fixed points). That the sphere $r = m/2$ is the fixed-point set of an isometry (I) implies that it is totally geodesic (has vanishing extrinsic curvature in Σ), as already discussed above. In particular it implies that $r = m/2$ is a minimal surface that joins two isometric halves. Hence (Σ, h) has two asymptotically flat ends, one for $r \mapsto \infty$ (spatial infinity) and one for $r \mapsto 0$, as shown on the left of Figure 12.7. This is sometimes interpreted by saying that there is a singular pointlike mass source at $r = 0$, just like for the electric Coulomb field for a point charge. But this interpretation is deceptive. It is true that the Coulomb field is a vacuum solution to Maxwell equations if the point at which the source sits is simply removed from space. But this removal of a point leaves a clear trace in that the resulting manifold is incomplete. This is different for the manifold $(\mathbb{R} - \{0\}, h)$, with h given by (12.29), which is complete, due to the fact that the origin is infinitely far away in the metric h . Hence no point is missing and the solution can be regarded as a genuine vacuum solution.

12.2.8 The $\mathbb{R}P^3$ Geon

There is a different twist to this story. One might object against the fact that $\mathbb{R} - \{0\}$ has *two* ends rather than just one (at spatial infinity). After all, what would the ‘inner end’ correspond to? A locally isometric manifold with just one end is obtained by taking the quotient of $\mathbb{R} - \{0\}$ with respect to the freely acting group \mathbb{Z}_2 that is

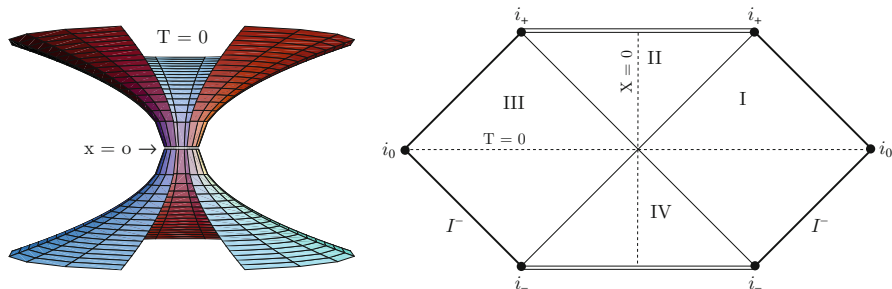


Fig. 12.7 To the right is the conformal (Penrose) diagram of Kruskal spacetime in which each point of this two-dimensional representation corresponds to a 2-sphere (an orbit of the symmetry group of spatial rotations). The asymptotic regions are i_0 (space-like infinity), I^\pm (future/past time-like infinity) and i^\pm (future/past time-like infinity). The diamond- and triangular-shaped regions I and II correspond to the exterior ($r > 2m$) and interior ($0 < r < 2m$) Schwarzschild spacetime, respectively, the interior being the black hole. The triangular region IV is the time reverse of II, a white hole. Region III is another asymptotically flat end isometric to the exterior Schwarzschild region I. The double horizontal lines at the top and bottom represent the singularities ($r = 0$) of the black and white hole, respectively. The left picture shows an embedding diagram of the hypersurface $T = 0$ (central horizontal line in the conformal diagram) that serves to visualise its geometry. Its minimal 2-sphere at the throat corresponds to the intersection of the hyperplanes $T = 0$ and $X = 0$ (bifurcate Killing Horizon)

generated by the isometry J in (12.31). This identifies the region $r > m/2$ with the region $r < m/2$ and antipodal points on the minimal 2-sphere $r > m/2$. The resulting space is real projective 3-space, $\mathbb{R}P^3$, minus a point, which clearly has just one end. Its full time evolution, i.e. the spacetime emerging from it, can be obtained from the maximal evolution of the Schwarzschild data: h as in (12.29), $K = 0$, which is Kruskal spacetime (see Kruskal 1960 and/or Chapter 5.5. in Hawking and Ellis (1973)). A conformally rescaled version (Penrose Diagram) of Kruskal spacetime is depicted on the right of Figure 12.7.

In Kruskal coordinates⁷ (T, X, θ, φ) , where T and X each range in $(-\infty, \infty)$ obeying $T^2 - X^2 < 1$, the Kruskal metric reads (as usual, we write $d\Omega^2$ for $d\theta^2 + \sin^2\theta d\varphi^2$)

$$g = \frac{32m^2}{\rho} e^{(-\rho/2m)} (-dT^2 + dX^2) + r^2 d\Omega^2, \tag{12.32}$$

where ρ is a function of T and X , implicitly defined by

$$((\rho/2m) - 1) e^{(\rho/2m)} = X^2 - T^2. \tag{12.33}$$

⁷In Kruskal (1960) Kruskal uses (v, u) ; Hawking and Ellis use (t', x') (Hawking and Ellis 1973) for what we call (T, X) .

Here ρ corresponds to the usual radial coordinate, in terms of which the Schwarzschild metric reads

$$g = - \left(1 - \frac{2m}{\rho} \right) dt^2 + \frac{dr^2}{1 - \frac{2m}{\rho}} + \rho^2 d\Omega^2 \quad (12.34)$$

where $\rho > 2m$. It covers region I of the Kruskal spacetime. Setting

$$\rho = r \left(1 + \frac{m}{2r} \right)^2 \quad (12.35)$$

so that the range $m/2 < r < \infty$ covers the range $2m < \rho < \infty$ twice, we obtain the ‘isotropic form’

$$g = - \left(\frac{1 - \frac{m}{2r}}{1 + \frac{m}{2r}} \right)^2 dt^2 + \left(1 + \frac{m}{2r} \right)^4 (dr^2 + r^2 d\Omega^2) \quad (12.36)$$

which covers regions I and III of the Kruskal manifold.

The Kruskal metric (12.32) is spherically symmetric and allows for the additional Killing field⁸:

$$K = \frac{1}{4m} (X\partial_T + T\partial_X), \quad (12.37)$$

which is time-like for $|X| > |T|$ and space-like for $|X| < |T|$.

The maximal time development of the \mathbb{RP}^3 initial data set is now obtained by making the following identification on the Kruskal manifold:

$$J : (T, X, \theta, \varphi) \mapsto (T, -X, \pi - \theta, \varphi + \pi). \quad (12.38)$$

It generates a freely acting group \mathbb{Z}_2 of smooth isometries which preserve space as well as time orientation. Hence the quotient is a smooth space- and time-orientable manifold, the \mathbb{RP}^3 *geon*.⁹ Its conformal diagram is just given by cutting away the $X < 0$ part (everything to the left of the vertical $X = 0$ line) in Figure 12.7 and

⁸That K is Killing is immediate, since (12.33) shows that ρ depends only on the combination $X^2 - T^2$ which is clearly annihilated by K .

⁹The \mathbb{RP}^3 geon is different from the two mutually different ‘elliptic interpretations’ of the Kruskal spacetime discussed in the literature by Rindler, Gibbons, and others. In Rindler (1965) the identification map considered is $J' : (T, X, \theta, \varphi) \mapsto (-T, -X, \theta, \varphi)$, which gives rise to singularities on the set of fixed-points (a 2-sphere) $T = X = 0$. Gibbons (1986) takes $J'' : (T, X, \theta, \varphi) \mapsto (-T, -X, \pi - \theta, \varphi + \pi)$, which is fixed-point free, preserves the Killing field (12.37) (which our map J does not), but does not preserve time-orientation. J'' was already considered in (Misner and Wheeler 1957, section 4.2), albeit in isotropic Schwarzschild coordinates already mentioned above, which only cover the exterior regions I and III of the Kruskal manifold.

taking into account that each point on the remaining edge, $X = 0$, now corresponds to a 2-sphere with antipodal identification, i.e. a \mathbb{RP}^2 (which is not orientable). The space-like hypersurface $T = 0$ has now the topology of the once punctured \mathbb{RP}^3 . In the left picture of Figure 12.7, this corresponds to cutting away the lower half and eliminating the inner boundary 2-sphere $X = 0$ by identifying antipodal points. The latter then becomes a minimal one-sided non-orientable surface in the orientable space-section of topology $\mathbb{RP}^3 - \{\text{point}\}$. The \mathbb{RP}^3 geon isometrically contains the exterior Schwarzschild spacetime (region I) with time-like Killing field K . But K ceases to exist globally on the geon spacetime since it reverses direction under (12.38).

12.3 X Without X

12.3.1 *Mass Without Mass*

At the end of Section 12.2.6, we already explained in what sense (active gravitational) mass emerges from pure topology and the constraints implied by Einstein's equation. Physically this just means that localised configurations of overall non-vanishing mass/energy may be formed from the gravitational field alone. With some care, one may say that such solutions represent bounded states of gravitons ('graviton balls'). However, they cannot be stable since gravitational solitons do not exist (in four spacetime dimensions)!

If Σ is topologically nontrivial, Gannon's theorem (Gannon 1975) (already discussed above) implies in full generality that the spacetime is singular (geodesically incomplete). The non-existence of vacuum, stationary, asymptotically flat spacetimes with non-vanishing mass, where the spacetime topology is $\mathbb{R} \times \Sigma$ and where Σ has only one end (spatial infinity), follows immediately from the expression (12.24) for the overall mass. Indeed, converting the surface integral (12.24) into a space integral via Stokes' theorem and using that $d \star dK^b$ is proportional to the spacetime's Ricci tensor shows¹⁰ that the expression vanishes identically due to the source-free Einstein equation. This generalises an older result due to Einstein and Pauli (1943) and is known as 'Lichnerowicz theorem', since Lichnerowicz first generalised the Einstein & Pauli result from static to stationary spacetimes, albeit using a far more involved argument than that given here (see Lichnerowicz (1955, livre premier, chapitre VIII)).

Most interestingly, this non-existence result ceases to be true in higher dimensions, as is exemplified by the existence of so-called Kaluza-Klein monopoles (Sorkin 1983) or (Gross and Perry 1983), which are nontrivial, regular, static and 'asymptotically flat' solutions to the source-free Einstein equations in a

¹⁰One uses the Killing identity $\nabla_a \nabla_b K_c = K_d R^d{}_{abc}$ to convert the second derivatives of K^b into terms involving no derivatives and the Riemann tensor.

five-dimensional spacetime. The crucial point to be observed here is that the Kaluza-Klein spacetime is ‘asymptotically flat’ in the sense that it is asymptotically flat in the ordinary sense for three spatial directions but not in the added fourth spatial direction, which is topologically a circle. Had asymptotic flatness in n dimensional spacetime been required for all $n - 1$ spatial directions, no such solution could exist (Deser 1988).

In this connection, it is interesting to note that in their paper, Einstein and Pauli (1943) actually claim to show the non-existence of soliton-like solutions in all higher-dimensional Kaluza-Klein theories even though they require asymptotic flatness in three spatial directions. But closer inspection reveals that their proof, albeit correct, invokes an additional and physically unjustified topological hypothesis that is violated by Kaluza-Klein monopoles. This is explained in more detail in Giulini (2008). Hence we may take Kaluza-Klein monopoles as a good example for the generation of mass and also magnetic charge in the framework of pure (higher dimensional!) general relativity without any sources.

12.3.2 *Momenta Without Momenta*

Source-free solutions with linear and angular momenta are also not difficult to obtain. Let us here just note a simple way of how to arrive at flux-integral expressions for these quantities. Let again (h, π) be a data set which is asymptotically flat on Σ with one end. Let ξ be a vector field on Σ that tends to the generator of an asymptotic isometry at infinity, that is, either a translation or a rotation. The corresponding linear or angular momentum is then just given by the usual *momentum map* corresponding to ξ :

$$\xi \mapsto \int_{\Sigma} d^3x \pi^{ab} L_{\xi} h_{ab} =: p_{\xi}, \quad (12.39)$$

where the right hand side is considered as a function on phase space $T^*\text{Riem}(\Sigma)$. Using the momentum constraint, $\nabla_a \pi^{ab} = 0$, an integration by parts in (12.39) converts it into a flux integral at spatial infinity which, re-expressing π in terms of K , reads

$$p_{\xi} := \lim_{R \rightarrow \infty} \left\{ \frac{1}{8\pi} \int_{S_R} (K_{ij} - h_{ij} h^{ab} K_{ab}) \xi^i n^j d\Omega \right\}. \quad (12.40)$$

This is the well-known expression for the ADM (Arnowitt, Deser, Misner) linear and angular momentum in geometric units.¹¹

¹¹That is, linear momentum has the unit of length (like mass) and angular momentum of length-squared. They are converted into ordinary units through multiplication with c/G .

Obviously there cannot exist a nontrivial asymptotically flat initial data set with an exact translational symmetry (because that translation could shift any local lump of curvature arbitrarily far into the asymptotically flat region, so that the curvature must be zero). But there may be such data sets with exact rotational symmetry. In that case, if ξ is the rotational Killing field and $\xi^b := g(\xi, \cdot)$ its associated one form in spacetime, a much simpler expression for angular momentum is given by the Komar integral (Komar 1959):

$$p_\xi := \lim_{R \rightarrow \infty} \left\{ \frac{1}{16\pi} \int_{S_R} \star d\xi^b \right\}, \quad (12.41)$$

where, as before, d is the exterior differential in spacetime and \star the Hodge dual with respect to the spacetime metric g . Again $d \star d\xi^b$ is proportional to the spacetime's Ricci tensor and hence zero, since spacetime is assumed to satisfy the source free Einstein equation. Hence Stokes' theorem implies that if Σ has only one end (spatial infinity) and the solution is regular in the interior, p_ξ must be zero. Therefore there cannot exist regular data set which give rise to rotationally symmetric solutions with non-vanishing angular momenta. A minimal relaxation is given by data sets which are *locally* symmetric, that is, in which a rotational Killing field exists up to sign. This slight topological generalisation indeed suffices to render the non-existence argument just given insufficient. Such data sets with net angular momentum have been constructed in Friedman and Mayer (1982).

12.3.3 Charge Without Charge

One case of 'charge without charge' is clearly given by the Kaluza-Klein monopoles mentioned above. Here we wish to stick to four spacetime dimensions and ask whether electric or magnetic charge can emerge from the Einstein-Maxwell equations without sources for the Maxwell field (in distinction to above, the energy-momentum tensor for the Maxwell field now acts as a source for the gravitational field).

If F is the 2-form on spacetime that represents the electromagnetic field, then the electric and magnetic charges q_e and q_m inside a 2-sphere S are, respectively, given by

$$q_e = \frac{1}{4\pi} \int_S \star dF, \quad (12.42a)$$

$$q_m = \frac{1}{4\pi} \int_S dF. \quad (12.42b)$$

Since $dF = 0$ (homogeneous Maxwell equation) and $d \star F = 0$ (inhomogeneous Maxwell equation with vanishing sources), these integrals depend only on the

homology class on S . This seems to imply that if spacetime has a regular interior, i.e. is of topology $\mathbb{R} \times \Sigma$ and Σ has only one end (spatial infinity), there will be no global net charge. The only possibility seems to be that there are local charges, e.g. if Σ has a wormhole topology $S^1 \times S^2 - \{\text{point}\}$, as shown by the lower-left drawing in Figure 12.6, where the flux lines thread through the wormhole. The homology class of 2-spheres that contain both wormhole mouths has zero charge, whereas the two individual wormhole mouths have equal and opposite charges associated with them.

However, this is not the only possibility! Our argument above relied on Stokes' theorem, which for ordinary forms presumes that the underlying manifold is orientable. On a non-orientable manifold, it only holds true for forms of density weight one, i.e. sections of the tensor bundle of forms twisted with the (now non trivial) orientation bundle (e.g. see Section 7 of Bott and Tu (1982)). This means that the argument for the non-existence of global charges can be extended to the non-orientable case for $\star F$ (which is a two form of density weight one) but not to F (which is a two form of density weight zero). Hence net electric charges cannot but magnetic¹² can exist (Sorkin 1977 and Friedman and Mayer (1982)). A simple illustration of how orientability comes into this is given by Figure 12.8.

As stated above, in the non-orientable case, Stokes theorem (here in three dimensions) continues to apply to two-forms of density weight one (e.g. the Hodge duals of one forms) and does not apply to two-forms of density weight zero, like the magnetic two form or, equivalently, its Hodge dual, which is a pseudo-vector field \mathbf{B} of zero divergence. We apply Stokes' theorem to suitable orientable submanifolds as explained in the caption to Figure 12.8. We obtain, denoting the flux of \mathbf{B} through a surface S with orientation O by $\Phi(\mathbf{B}, S, O)$,

$$\Phi(\mathbf{B}, \partial\Sigma_1, O) + \Phi(\mathbf{B}, S_1, O) + \Phi(\mathbf{B}, S_2, O) = 0 \tag{12.43}$$

in the first case and

$$\Phi(\mathbf{B}, S_1, O') + \Phi(\mathbf{B}, S_2, O) = 0 \tag{12.44}$$

in the second. Using the obvious fact that the flux integral changes sign if the orientation is reversed, i.e. $\Phi(\mathbf{B}, S_1, O') = -\Phi(\mathbf{B}, S_1, O)$, we get

$$\Phi(\mathbf{B}, \partial\Sigma_1, O) = -2\Phi(\mathbf{B}, S_1, O). \tag{12.45}$$

¹²The distinction between electric and magnetic is conventional in Einstein-Maxwell theory without sources for the Maxwell field, since the energy-momentum tensor T for the latter is invariant under duality rotations which rotate between F and $\star F$ according to $\omega \mapsto e(i\varphi)\omega$, where $\omega := F + i\star F$. Since $T_{\mu\nu} \propto \omega_{\mu\lambda}\bar{\omega}_\nu{}^\lambda$, where an overbar denotes complex conjugation, the invariance of T is immediate.

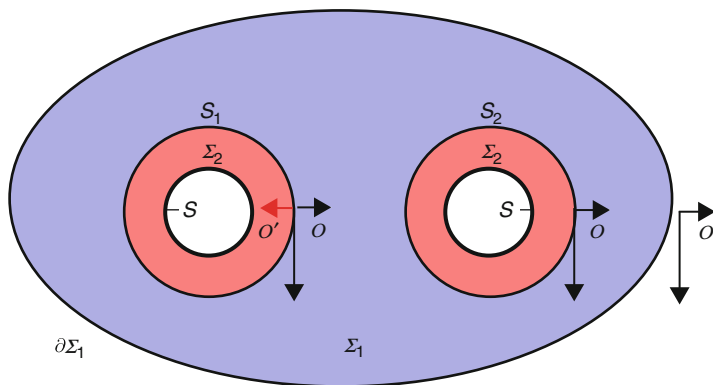


Fig. 12.8 Consider the three-dimensional region that one obtains by rotating this figure about the central horizontal axis of symmetry. The two inner boundary spheres S are to be identified in a way so that their induced orientations O match, e.g. by simple translation. (In this two-dimensional picture, an orientation is represented by an ordered two-leg, where the ordering is according to the different lengths of the legs.) This results in a *non-orientable* manifold with single outer boundary component $\partial\Sigma_1$, corresponding to the non-orientable wormhole. In the text, we apply Stokes' theorem twice to two orientable submanifolds: first, to the heavier shaded region bounded by the outer 2-sphere $\partial\Sigma_1$ with orientation O and the inner two 2-spheres S_1 and S_2 with like orientations O as indicated and second to the lightly shaded cylindrical region labelled by Σ_2 that is bounded by the two 2-spheres S_1 and S_2 with opposite orientations O' and O , respectively

So in order to get a nonzero global charge, we just need to find a divergenceless pseudo-vector field on Σ_1 with non-vanishing flux through S , which can be arranged.

Note that the trick played here in using non-orientable Σ would not work for the Komar integrals (12.24) (12.41), since the Hodge map \star turns the ordinary two-form dK^b (or $d\xi^b$) of density weight zero into the two-form $\star dK^b$ (or $\star d\xi^b$) of density weight one, so that Stokes' theorem continues to hold in these cases for non-orientable Σ by the result cited above.

12.3.4 Spin Without Spin

In my opinion, by far the most surprising case of 'X without X' is that where X stands for *spin*, i.e. half-integral angular momentum. It was certainly not anticipated by Misner, Thorne and Wheeler, who in their otherwise most comprehensive book (Misner et al. 1973) were quite lost in trying to answer their own question of how 'to find a natural place for spin 1/2 in Einstein's standard geometrodynamics' (Misner et al. 1973, Box 44.3). A surprising answer was offered 8 years later, in 1980, by John Friedman and Rafael Sorkin (1980).

It is often said that the need to go from the group $SO(3)$ of spatial rotations to its double (= universal) cover, $SU(2)$, is quantum-mechanical in origin and cannot be

understood on a classical basis. In some sense, the mathematical facts underlying the idea of ‘spin 1/2 from gravity’ disprove this statement. They imply that if the 3-manifold Σ has a certain topological characteristic, the asymptotic symmetry group for isolated systems (modelled by spatially asymptotically flat data) is not the Poincaré group (inhomogeneous Lorentz group) in the sense of Beig and Murchadha (1987), but rather its double (= universal) covers – for purely topological reasons! Let us try to explain all this in more detail.

Recall that in quantum mechanics, the possibility for this enlargement (central extension) of a classical symmetry group has its origin in the assumption that the phase of the complex wave function is a redundant piece of description (i.e. unobservable), at least for states describing isolated systems, so that symmetry groups should merely act on the space of rays rather than on Hilbert space by proper representations. Hence it is sufficient for the symmetry group to be implemented by so-called ray representations, which in case of the rotation group are in bijective correspondence to proper representations of its double (= universal) cover group. Accordingly, in quantum mechanics, there exist physically relevant systems whose state spaces support proper representations of $SU(2)$ but not of $SO(3)$: These are just the systems whose angular momentum is an odd multiple of $\hbar/2$. We will say that such systems admit *spinorial states*.

Spinorial states are not necessarily tight to the usage of spinors. They also have a place in ordinary Schrödinger quantisation, i.e. for systems whose quantum state space is represented by the Hilbert space of square integrable functions over the classical configuration space Q , which here and in what follows is understood to be the reduced configuration space in case constraints existed initially. Then, spinorial states exist if the following conditions hold:

- S1 Q is not simply connected.
- S2 The (say left) action $SO(3) \times Q \rightarrow Q$, $(g, q) \mapsto g \cdot q$, of the ordinary rotation group $SO(3)$ on the classical configuration space Q is such that if $\gamma : [0, 2\pi] \mapsto SO(3)$ is any full 360-degree rotation about some axis, then the loop $\Gamma := \gamma \cdot q$ in Q , based at $q \in Q$, is not contractible, i.e. defines a nontrivial element (of order two since γ traversed twice is contractible in $SO(3)$) in $\pi_1(Q, q)$, the fundamental group of Q based at q . It is not hard to see that this property (of being non-contractible) is independent of the basepoint $q \in Q$ within the same path component of Q , though it may vary if one goes from one path component to another (as in the Skyrme model mentioned below). See Giulini (1993), in particular the proof of Lemma 1.

The reason for the existence of spinorial states in such situations lies in possible generalisations of Schrödinger quantisation if the domain for the wave function is a space, Q , whose fundamental group is nontrivial. The idea of generalisation is to define the Schrödinger function on the universal cover space \tilde{Q} (i.e. the Hilbert space is the space of square-integrable functions on \tilde{Q}) but to restrict the observables to those that commute with the unitary action of the deck transformations. The latter then form a discrete gauge group isomorphic to the fundamental group of Q . The Hilbert space decomposes into superselection sectors which are labelled by

the equivalence classes of unitary irreducible representations. The sector labelled by the trivial class is isomorphic to that of ordinary Schrödinger quantisation on Q , whereas the other sectors are acquired through the generalisation discussed here.

This is related to, but not identical with, another generalisation that is usually mentioned in the context of geometric quantisation. There one generalises Schrödinger quantisation by considering quantum states as square-integrable *sections* in a complex line bundle over Q (rather than just complex-valued functions on Q). This leads to additional sectors labelled by the equivalence classes of complex line bundles, which are classified by $H^2(Q, \mathbb{Z})$, the second cohomology group of Q with integer coefficients (see, e.g. Woodhouse 1991).

In generalised Schrödinger quantisation spinorial states will correspond to particular such new sectors. To make this more precise in the geometric-quantisation picture, we recall that $H^2(Q, \mathbb{Z})$, being a finitely generated abelian group, has the structure

$$H^2(Q, \mathbb{Z}) \cong \underbrace{\mathbb{Z} \oplus \cdots \oplus \mathbb{Z}}_{\text{free part}} \oplus \underbrace{\mathbb{Z}_{p_1} \oplus \cdots \oplus \mathbb{Z}_{p_n}}_{\text{torsion part}} . \tag{12.46}$$

The number of factors \mathbb{Z} in the free part is called the second Betti number and the number n of cyclic groups the second torsion number. For this to be well defined, we have to agree that each of the integers p_i should be a power of a prime.¹³ Spinorial states are then given by sections in all those line bundles which represent *a particular* \mathbb{Z}_2 factor in the torsion part of its decomposition according to (12.46) non-trivially.

We said ‘a particular \mathbb{Z}_2 factor’. Which one? The answer is: That one, which is generated by the 360-degree rotation according to criterion S2 above. To understand this, we remark that the torsion part of $H^2(Q, \mathbb{Z})$ can be understood in terms of the fundamental group. More precisely, the torsion part of $H^2(Q, \mathbb{Z})$ is isomorphic to the torsion part of the abelianisation of the fundamental group.¹⁴ Given that isomorphism, we can now identify the \mathbb{Z}_2 factor in $H^2(Q, \mathbb{Z})$ with the \mathbb{Z}_2 subgroup of the fundamental group that is generated by 360-degree rotations, as explained by S2.

A simple illustrative example of this is given by the rigid rotor, that is, a system whose configuration space Q is the group manifold $SO(3)$, which as manifold is isomorphic to $\mathbb{R}P^3$. The action of physical rotations is then given by left translation. Here we have $H^2(Q, \mathbb{Z}) \cong \mathbb{Z}_2$, i.e. it is pure torsion and, in fact, isomorphic to the fundamental group. Quantisation then leads to two sectors: Those containing states

¹³A classic theorem on finite abelian groups states that if p and q are integers, \mathbb{Z}_{pq} is isomorphic to $\mathbb{Z}_p \oplus \mathbb{Z}_q$ iff p and q are coprime.

¹⁴This follows in two steps: First, one recalls $H^2(Q, \mathbb{Z})$ is isomorphic to the direct sum of the free part of $H_2(Q, \mathbb{Z})$ and the torsion part of $H_1(Q, \mathbb{Z})$ (universal coefficient theorem). Second, one uses that $H_1(Q, \mathbb{Z})$ is isomorphic to the abelianisation of the fundamental group (Hurewicz’ theorem).

of integral spin, which are represented by ordinary square integrable functions on Q , and those containing half-integral spin, represented by square integrable sections in the unique nontrivial line bundle over $Q \cong \mathbb{RP}^3$.

More sophisticated field theoretic examples for this mechanism are given by so-called non-linear sigma models, in which the physical states are given by maps from physical space into some non-linear space of field values, like, e.g. a sphere. A particular such model is the Skyrme model (Skyrme 1971), in which the target space is the three-sphere S^3 . Configurations of finite energy must map spatial infinity (physical space is \mathbb{R}^3) into a single point of S^3 so that Q decomposes into a countably infinity of path components according to the winding number of that map. In the Skyrme model, which serves to give an effective description of baryons, this winding number corresponds to the baryon number. The fundamental group of each path component is isomorphic to the fourth homotopy group of the target space S^3 , which is again just \mathbb{Z}_2 . One can now prove that the loops traced through by 360-degree rotations are contractible in the components of even winding numbers and non-contractible in the components of odd winding numbers (Giulini 1993). Hence spinorial states exist for odd baryon numbers, as one should expect on physical grounds.

These examples differ from those in general relativity insofar as in the latter spinorial states usually exist only in non-abelian sectors, i.e. sectors that correspond to higher-dimensional unitary irreducible representations of the fundamental group (Giulini 1995a). An example will be mentioned below. For that reason, we made the distinction between the first and the second (geometric quantisation) methods of generalising Schrödinger quantisation, since non-abelian sectors are obtained in the first but not in the second method, which is only sensitive to the abelianisation of the fundamental group. That the restriction to abelian sectors is unnecessary and unwarranted is further discussed in Giulini (1995b).

The geometric-topological situation underlying the existence of spinorial states in general relativity is this (Friedman and Sorkin 1980): Consider a 3-manifold Σ with one regular end, so as to describe an asymptotically flat isolated system without internal infinities. Here ‘regular’ means that the one-point compactification $\bar{\Sigma}$ of Σ is again a manifold. This means that Σ contains a compact subset the complement of which is a cylinder $\mathbb{R} \times S^2$. A physical rotation of the system so represented is then given by a diffeomorphism whose support is entirely on that cylinder and rotates the S^2 at one end relative to the S^2 at the other end by full 360 degrees (see Figure 12.9).

The question is now this: Is this diffeomorphism in the identity component of those diffeomorphisms that fix the 2-sphere at ‘spatial infinity’? See Figure 12.10 for further illustration. The answer to this question just depends on the topology on Σ and is now known for all 3 manifolds.¹⁵ Roughly speaking, the generic case is that

¹⁵To decide this entails some subtle issues, like whether to diffeomorphisms that are homotopic (continuously connected through a one-parameter family of *continuous* maps) are also isotopic (continuously connected through a one-parameter family of *homeomorphisms*) and then also

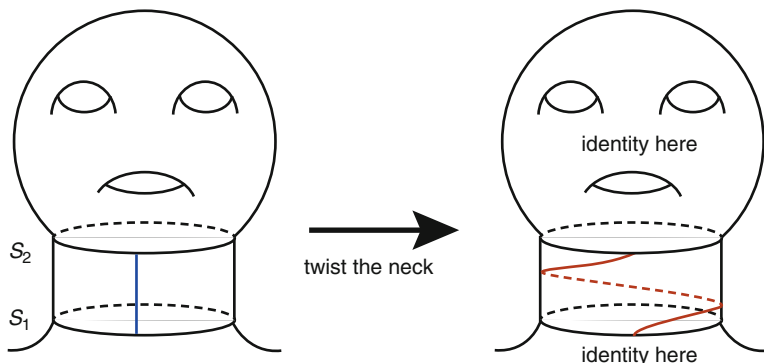


Fig. 12.9 A full 360-degree rotation of the part of the manifold above the 2-sphere S_2 relative to the part below the 2-sphere S_1 is given by a diffeomorphism with support on the cylinder region bounded by S_1 and S_2 that rotates one bounding sphere relative to the other by 360 degrees ('twisting the neck' by 360 degrees)

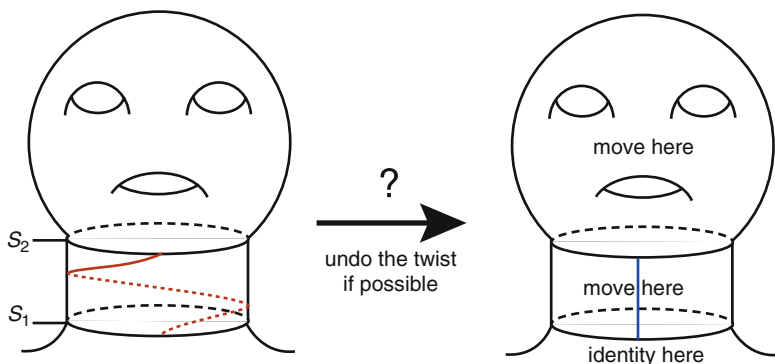


Fig. 12.10 Keeping all points below S_1 fixed but now allowing the points above S_2 to move, the neck-twist may or may not be continuously undone through a continuous sequence of diffeomorphisms whose support is entirely above the sphere S_1 . If it cannot be undone in this fashion, the manifold above S_2 , or rather its one-point compactification, is called 'spinorial'. Do not be misled to think that you can just undo it by rigidly rotating the upper part in the embedding space shown here, since this will generally not define a diffeomorphism of the manifold itself

diffeotopic (continuously connected through a one-parameter family of *diffeomorphisms*). The crucial question is whether homotopy implies isotopy, which is not at all obvious since on a homotopy the interpolating maps connecting two diffeomorphisms are just required to be continuous, that is, they need not be continuously invertible as for an isotopy. For example, the inversion $I(\mathbf{x}) = -\mathbf{x}$ in \mathbb{R}^n is clearly not isotopic to the identity, but homotopic to it via $\phi_t(\mathbf{x}) = (1 - 2t)\mathbf{x}$ for $t \in [0, 1]$. Then $\phi_0 = \text{id}$, $\phi_1 = I$ and only at $t = 1/2$ does the map ϕ_t cease to be invertible.

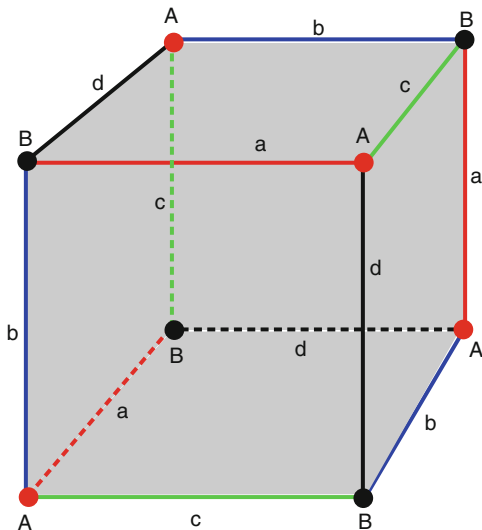
spinorial spaces are allowed. More precisely, those 3-manifold $\bar{\Sigma}$ (from now on we represent the manifolds by their one-point compactifications in order to talk about closed spaces) which do not allow for spinorial states are connected sums of lens spaces and handles ($S^1 \times S^2$). This is a very nice (though rather nontrivial) result insofar, as the non-spinoriality of these spaces as well as their connected sums is easy to visualise. Hence one may say that there are no other non-spinorial manifolds than the ‘obvious’ ones.

Take, for example, the simplest lens space¹⁶ $L(2, 1)$, which is just real projective 3-space $\mathbb{R}P^3$. It can be imagined as a solid ball B in \mathbb{R}^3 with antipodal points on the 2-sphere boundary identified. Think of an inner point, say the centre, of B as the point at infinity, and surround it by a small spherical shell whose inner boundary is the 2-sphere S_1 and outer boundary the 2-sphere S_2 (above we called it a ‘cylinder’ since its topology is $\mathbb{R} \times S^2$). Now perform a full 360-degree rotation of S_1 against S_2 with support inside the shell. Can this diffeomorphism be undone through a continuous sequence of diffeomorphisms that fix all points inside the inner sphere S_1 ? Clearly it can: Just rigidly rotate the outside to undo it. The crucial point is that this rigid rotation is compatible with the boundary identification and hence does indeed define a diffeomorphism of $\mathbb{R}P^3$. Essentially the same argument applies to all other ‘obvious cases’.

In contrast, it is much more difficult to prove that such an undoing is impossible, i.e. the spinoriality of a given manifold. Needless to say, the fact that you cannot easily visualise a possible undoing of a 360 degree twist does not mean it does not exist. A simple and instructive example is given by the spherical space form S^3/D_8^* , where D_8^* is the eight-element non-abelian subgroup of $SU(2)$ that doubly covers (via the double cover $SU(2) \rightarrow SO(3)$) the four-element abelian subgroup of $SO(3)$ that is given by the identity and the three 180-degree rotations about the mutually perpendicular x , y , and z axes. Identifying S^3 with $SU(2)$ the quotient S^3/D_8^* is defined by letting D_8^* act through, say, right translations. Since $SU(2)$ is also the group of unit quaternions, D_8^* can be identified with its subgroup $\{\pm 1, \pm i, \pm j, \pm k\}$, where i, j, k denote the usual unit quaternions (they square to -1 and $ij = k$ and also cyclic permutations thereof). A way to visualise S^3/D_8^* is given in Figure 12.11. Note that if the 2-dimensional boundary of the cube is smoothly deformed to a round 2-sphere, a rigid rotation in the embedding \mathbb{R}^3 would still not be compatible with the boundary identifications. In fact, it is known that S^3/D_8^* is spinorial (see (Giulini 1994) for more information and

¹⁶ The definition of lens spaces $L(p, q)$ in three dimensions is $L(p, q) = S^3/\sim$, where (p, q) is a pair of positive coprime integers with $p > 1$, $S^3 = \{(z_1, z_2) \in \mathbb{C}^2 \mid |z_1|^2 + |z_2|^2 = 1\}$, and $(z_1, z_2) \sim (z'_1, z'_2) \Leftrightarrow z'_1 = e(2\pi i/p)z_1$, and $z'_2 = e(2\pi i q/p)z_2$. One way to picture the space is to take a solid ball in \mathbb{R}^3 and identify each points on the upper hemisphere with points on the lower hemisphere after a rotation by $2\pi q/p$ about the vertical symmetry axis. In this way, each set of p equidistant points on the equator is identified to a single point. The fundamental group of $L(p, q)$ is \mathbb{Z}_p , i.e. independent of q , and the higher homotopy groups are those of its universal cover, S^3 . This does, however, not imply that $L(p, q)$ is homotopy equivalent, or even homeomorphic, to $L(p, q)$. The precise relation will be stated below.

Fig. 12.11 The manifold S^3/D_8^* is obtained from a solid cube by identifying opposite faces after a relative 90-degree rotation about the axis connecting their midpoints. In the picture shown here the identifying motion between opposite faces is a right screw, giving rise to the identifications of edges and vertices as labelled in the picture



references). Here we just remark that D_8^* has five equivalence classes of unitary irreducible representations: four one-dimensional and a single two-dimensional one. The spinorial sector is that corresponding to the latter, that is, it is a non-abelian sector.

Another remarkable property of S^3/D_8^* is that it is *chiral*, that is, it does not admit for orientation reversing self-diffeomorphisms (see Giulini 1994 for information about which 3-manifolds are chiral and Muellner 2008 for a recent systematic investigation of chirality in all dimensions). This means that if we had chosen the map that identifies opposite faces of the cube shown in Figure 12.11 to be a left rather than right screw, we would have obtained a manifold that is not orientation-preserving diffeomorphic to the one originally obtained, though they are clearly orientation-reversing diffeomorphic as they are related by a simple reflection at the origin of the embedding \mathbb{R}^3 . Being chiral seems to be more the rule than the exception for 3-manifolds (Giulini 1994).

12.4 Further Developments

In the last subsection, we have learnt that the fundamental group of the configuration space of the gravitational field will give rise to sectors with potentially interesting physical interpretations. Hence it seems natural to generally ask: What is the fundamental group of the configuration space associated to a manifold Σ ? The last question can be given an elegant abstract answer, though not one that will always allow an easy characterisation (determination of the isomorphism class) of the group. The abstract answer is in terms of a presentation of a certain mapping-class

group and comes about as follows: Consider the 3-manifold Σ which we assume to have one regular end. Hence its one-point compactification, $\bar{\Sigma}$, is a manifold. Next consider the group of diffeomorphisms $\text{Diff}_F(\Sigma)$ that fix a prescribed point $p \in \bar{\Sigma}$ as well as all vectors in the tangent space at this point. It is useful to think of p as the ‘point at infinity’, i.e. the point that we added for compactification, for then it is intuitively clear that $\text{Diff}_F(\bar{\Sigma})$ corresponds to those diffeomorphism of Σ that tend to the identity as one moves to infinity within the single end. In order to have that picture in mind, we will from now on write ∞ for the added point p . The configuration space of the gravitational field on Σ can then be identified with the space of Riemannian metrics on $\bar{\Sigma}$, $\text{Riem}(\bar{\Sigma})$, modulo the identifications induced by $\text{Diff}_F(\bar{\Sigma})$, i.e.

$$Q(\Sigma) = \text{Riem}(\bar{\Sigma})/\text{Diff}_F(\bar{\Sigma}) . \tag{12.47}$$

Now, it is true that $\text{Diff}_F(\bar{\Sigma})$ acts freely on $\text{Riem}(\bar{\Sigma})$ (there are no nontrivial isometries on a Riemannian manifold that fix a point and the frame at that point) and that this action admits a slice (see Ebin 1968). Hence $\text{Riem}(\bar{\Sigma})$ is a principle fibre bundle with group $\text{Diff}_F(\bar{\Sigma})$ and base $Q(\Sigma)$ (Fischer 1970, 1986). But $\text{Riem}(\bar{\Sigma})$ is contractible (being an open positive convex cone in the vector space of smooth sections of symmetric tensor fields of rank two over $\bar{\Sigma}$). Hence the long exact-sequence of homotopy groups for the fibration $\text{Diff}_F(\bar{\Sigma}) \rightarrow \text{Riem}(\bar{\Sigma}) \rightarrow Q(\Sigma)$ implies the isomorphism of the n th homotopy group of the fibre $\text{Diff}_F(\bar{\Sigma})$ with the $n + 1$ st homotopy group of the base $Q(\Sigma)$. In particular, the first homotopy group (i.e. the fundamental group) of $Q(\Sigma)$ is isomorphic to the zeroth homotopy group of the group $\text{Diff}_F(\bar{\Sigma})$. However, the latter is just the quotient $\text{Diff}_F(\bar{\Sigma})/\text{Diff}_F^0(\bar{\Sigma})$, where $\text{Diff}_F^0(\bar{\Sigma}) \subset \text{Diff}_F(\bar{\Sigma})$ is the normal subgroup formed by the connected component of $\text{Diff}_F(\bar{\Sigma})$ that contains the identity. In this way, we finally arrive at the result that the fundamental group of $Q(\Sigma)$ is isomorphic to a *mapping-class* group:

$$\pi_1(Q(\Sigma)) \cong \text{Diff}_F(\bar{\Sigma})/\text{Diff}_F^0(\bar{\Sigma}) . \tag{12.48}$$

This is a very interesting result in its own right. It contains the mathematical challenge to characterise $\text{Diff}_F(\bar{\Sigma})/\text{Diff}_F^0(\bar{\Sigma})$. A way to attack this problem is to use the fact that each element in $\text{Diff}_F(\bar{\Sigma})$ defines a self-map $\pi_1(\bar{\Sigma}, \infty)$ just by mapping loops based at ∞ to their image loops, which are again based at ∞ since elements of $\text{Diff}_F(\bar{\Sigma})$ keep that point fixed. Since in this fashion homotopic loops are mapped to homotopic loops, this defines indeed a map on $\pi_1(\bar{\Sigma}, \infty)$ which is, in fact, an automorphism. Moreover, elements in the identity component $\text{Diff}_F^0(\bar{\Sigma})$ give rise to the trivial automorphisms. This is obvious, since the images of a loop under continuously related diffeomorphisms will in particular result in homotopic loops. Hence we have in fact a homomorphism from $\text{Diff}_F(\bar{\Sigma})/\text{Diff}_F^0(\bar{\Sigma})$ into the automorphism group of $\pi_1(\bar{\Sigma}, \infty)$:

$$h : \text{Diff}_F(\bar{\Sigma})/\text{Diff}_F^0(\bar{\Sigma}) \rightarrow \text{Aut}(\pi_1(\bar{\Sigma}, \infty)). \tag{12.49}$$

The strategy is now this: Assume we know a presentation of $\text{Aut}(\pi_1(\bar{\Sigma}, \infty))$, that is, a characterisation of this group in terms of generators and relations. Then we aim to make useful statements about the kernel and image of the map in (12.49) so as to be able to derive a presentation for $\text{Diff}_F(\bar{\Sigma})/\text{Diff}_F^0(\bar{\Sigma})$. A simple but nontrivial example will be given below. We recall that $\bar{\Sigma}$ is a unique connected sum of prime manifolds and that $\pi_1(\bar{\Sigma})$ is the free product of the fundamental groups of the primes. Since (finite) presentations for the automorphism group of a free product can be derived if (finite) presentations for the automorphism groups of the factors are known (Gilbert 1987), we in principle only need to know the latter.

Another mathematically interesting aspect connected with (12.48) is the fact that $\text{Diff}_F(\bar{\Sigma})/\text{Diff}_F^0(\bar{\Sigma})$ is a topological invariant of $\bar{\Sigma}$ which is *not* a homotopy invariant (McCarty and Shultz 1963). Hence (12.48) implies that $\pi_1(Q(\Sigma))$, too, is a topological invariant of $\bar{\Sigma}$ which is not homotopy invariant, i.e. it might tell apart 3-manifolds which are homotopically equivalent but not homeomorphic. There are indeed examples for this to happen. Here is one: Recall that lens spaces (see footnote 16) $L(p, q)$ and $L(p, q')$ are homotopy equivalent iff $qq' = \pm n^2 \pmod p$ for some integer n (Whitehead 1941, theorem 10) and homeomorphic¹⁷ iff (all four possibilities) $q' = \pm q^{\pm 1} \pmod p$ (here all four possibilities of combinations of \pm signs are considered). On the other hand, the mapping-class group $\text{Diff}_F(\bar{\Sigma})/\text{Diff}_F^0(\bar{\Sigma})$ for $L(p, q)$ is $\mathbb{Z} \times \mathbb{Z}$ if $q^2 = 1 \pmod p$ with $q \neq \pm 1 \pmod p$ and \mathbb{Z} in the remaining cases for $p > 2$ (see Table IV on p. 591 of Witt 1986). Take now, as an example, $p = 15, q = 1$, and $q' = 4$. Then the foregoing implies that $L(15, 1)$ and $L(15, 4)$ are homotopic but not homeomorphic and have different mapping-class groups.

Finally we give an example for a presentation and its pseudo-physical interpretation for $\text{Diff}_F(\bar{\Sigma})/\text{Diff}_F^0(\bar{\Sigma})$. Consider the connected sum (denoted by #) of two real projective spaces $\mathbb{R}P^3$. This manifold may be visualised as explained in Figure 12.12.

The fundamental group of $\mathbb{R}P^3\#\mathbb{R}P^3$ is the twofold free product $\mathbb{Z}_2 * \mathbb{Z}_2$ of the fundamental group \mathbb{Z}_2 of the factors $\mathbb{R}P^3$. With respect to the generators a and b shown in Figure 12.13 or, alternatively, with respect to the generators a and c , where $c := ab$ corresponds to the loop shown by the two horizontal segments in Figure 12.12, two alternative presentations of the fundamental group are given by

$$\pi_1(\mathbb{R}P^3\#\mathbb{R}P^3) = \underbrace{\langle a, b \mid a^2 = b^2 = 1 \rangle}_{\cong \mathbb{Z}_2 \times \mathbb{Z}_2} = \underbrace{\langle a, c \mid a^2 = 1, aca^{-1} = c^{-1} \rangle}_{\cong \mathbb{Z}_2 \rtimes \mathbb{Z}} \tag{12.50}$$

¹⁷As regards the notion of chirality, an interesting refinement of this statement is that $L(p, q)$ and $L(p, q')$ are *orientation-preserving* homeomorphic iff $q' = q^{\pm 1} \pmod p$ (Reidemeister 1935).

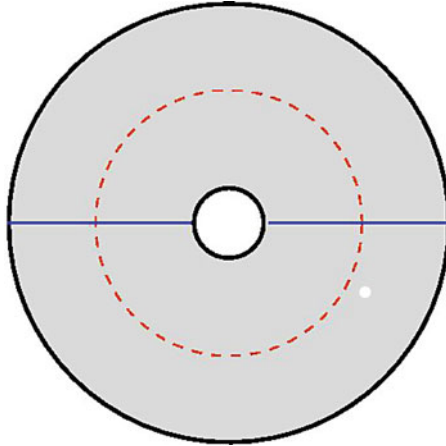


Fig. 12.12 The connected sum $\mathbb{R}P^3 \# \mathbb{R}P^3$ between two real projective spaces may be visualised as a spherical shell (here the grey-shaded region) where antipodal points on each of the two 2-sphere boundaries, S_1 and S_2 , are identified. The two-dimensional figure here should be rotated about the horizontal symmetry axis. The two horizontal line segments shown form a circle in view of the antipodal identifications. It shows that $\mathbb{R}P^3 \# \mathbb{R}P^3$ is a circle bundle over $\mathbb{R}P^2$. The dotted circle, which upon rotation of the figure becomes a 2-sphere, can be thought of as the 2-sphere along which the connected sum between the two individual $\mathbb{R}P^3$ manifolds is taken

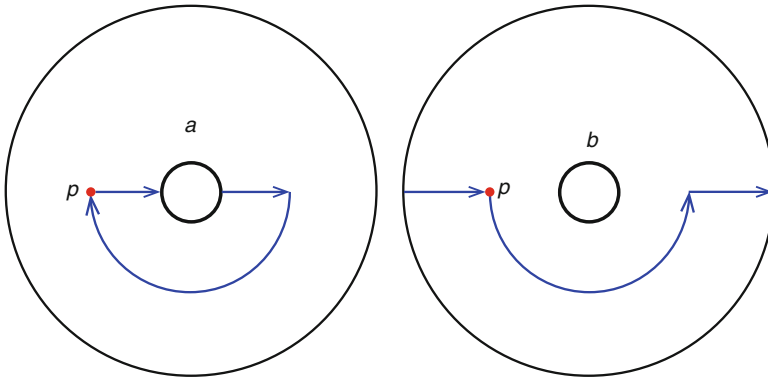


Fig. 12.13 Shown are two closed loops based at some point p whose homotopy classes generate the fundamental group $\mathbb{Z}_2 * \mathbb{Z}_2 \cong \mathbb{Z}_2 \times \mathbb{Z}$ of the manifold $\mathbb{R}P^3 \# \mathbb{R}P^3$

Turning to (12.49), we first remark that the automorphism group of $\mathbb{Z}_2 * \mathbb{Z}_2$ is itself isomorphic to $\mathbb{Z}_2 * \mathbb{Z}_2$,

$$\text{Aut}(\mathbb{Z}_2 * \mathbb{Z}_2) \cong \mathbb{Z}_2 * \mathbb{Z}_2 = \langle E, S \mid E^2 = S^2 = 1 \rangle, \tag{12.51}$$

where the two generators E and S can be identified by stating their action on the generators a and b of the fundamental group:

$$E : (a, b) \mapsto (b, a), \quad S : (a, b) \mapsto (a, aba^{-1}). \quad (12.52)$$

It may now be shown that the map h in (12.49) is an isomorphism so that the fundamental group of the configuration space $Q(\Sigma)$ is the free product $\mathbb{Z}_2 * \mathbb{Z}_2$. Injectivity of h is not so obvious (but true) whereas surjectivity can be shown by visualising diffeomorphisms that actually realise the generators E and S of (12.52). For example, E can be realised by an inversion on the sphere along which the connected sum is taken (see Figure 12.12) (which is orientation reversing) followed by a simple reflection along a symmetry plane (so as to restore orientation preservation). Its ‘physical’ meaning is that of an exchange of the two diffeomorphic factors (primes) in the connected-sum decomposition. The map for S is a little harder to visualise since it mixes points between the two factors (see Giulini 2007 for pictures). It can roughly be described as sliding one factor through the other and back to its original position. Here we wish to focus on the following: Given the generalisations of Schrödinger quantisations outlined above, we are naturally interested in the equivalence classes of unitary irreducible representations of $\mathbb{Z}_2 * \mathbb{Z}_2$. They can be obtained by elementary means and are represented by the four obvious one-dimensional representations where $E \mapsto \pm 1$ and $S \mapsto \pm 1$ and a continuum of two-dimensional ones where

$$E \mapsto \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad S \mapsto \begin{pmatrix} \cos \tau & \sin \tau \\ \sin \tau & -\cos \tau \end{pmatrix}, \quad \tau \in (0, \pi). \quad (12.53)$$

No higher dimensional ones occur. The one-dimensional representations already show that both statistics sectors exist. Moreover, the two-dimensional representations show that the diffeomorphisms representing S mix the statistics sectors by an angle τ that depends on the representation class. All this may be read as indication against a classical ‘spin-statistics correlation’ that one might have expected from experience with other non-linear field theories, e.g. following Finkelstein and Rubinstein (1968) and Sorkin (1988). Such a connection can therefore only exist in certain sectors and the question can (and has) be asked how these sectors are selected (Dowker and Sorkin 1998, 2000). See Giulini (1994, 1997) for other examples with explicit presentations of $\text{Diff}_{\mathbb{F}}(\bar{\Sigma})/\text{Diff}_{\mathbb{F}}^0(\bar{\Sigma})$ where $\bar{\Sigma}$ is either the n fold connected sum of real projective spaces $\mathbb{R}P^3$ or handles $S^1 \times S^2$ and also some general statements.

From what has been said so far, it clearly emerges that the enormous topological variety and complexity of 3-manifolds leave their structural traces in general relativity, which can be used to model some of the properties in pure gravity that are usually associated with ordinary matter. This is indeed made practical use of, e.g. in modelling scattering and merging processes of black holes with data corresponding to wormhole topologies. But one should also say that the physical relevance of much of what I said later is not at all established. The aim of my presentation was to

alert to the existence of these structures, leaving their physical relevance open for the time being. Somehow all this may remind one Tait's beautiful idea to model the discrete structural properties of material atoms on the properties of knots in physical space, which he thought of as knotted vortex lines in the all-embracing hypothetical ether medium. But whereas there was never formulated a fundamental dynamical theory of the ether,¹⁸ there is a well-formulated and well-tested theory of geometrodynamics: General Relativity. In that sense, we are in a much better position than Tait was in the mid-1880s.

Acknowledgements I sincerely thank the organisers of the *Beyond Einstein* conference at Mainz University for inviting me to this most stimulating and pleasant meeting.

References

- Arnowitt, R., Deser, S., & Misner C. W. (1962). The dynamics of general relativity. In L. Witten (Ed.), *Gravitation: An introduction to current research* (pp. 227–265). New York and London: Wiley.
- Beig, R., & Ó Murchadha, N. (1987). The Poincaré group as symmetry group of canonical general relativity. *Annals of Physics*, 174, 463–498.
- Bott, R., & Tu, L. W. (1982). *Differential forms in algebraic topology*. Graduate texts in mathematics. New York: Springer.
- Clifford, W. K. (1982). *Mathematical Papers* (1st ed.). London: Macmillan (1882). R. Tucker (Ed.)
- Deser, S. (1988). Absence of regular static Einstein solutions in arbitrary dimensions. *Classical and Quantum Gravity*, 5(1), L9–L10.
- DeWitt, B. (1967). Quantum theory of gravity I. *Physical Review*, 160, 1113–1148.
- Dowker, F., & Sorkin, R. (1998). A spin-statistics theorem for certain topological geons. *Classical and Quantum Gravity*, 15, 1153–1167.
- Dowker, F. & Sorkin, R. (2000). Spin and statistics in quantum gravity. In R. C. Hilborn & G. M. Tino (Eds.), *Spin-statistics connections and commutation relations: experimental tests and theoretical implications* (pp. 205–218). New York: American Institute of Physics.
- Ebin, D. G. (1968). On the space of Riemannian metrics. *Bulletin of the American Mathematical Society*, 74(5), 1001–1003.
- Einstein, A. (1920). Äther und Relativitätstheorie. Reprinted in M. Janssen, et al. (Eds.), *The collected papers of Albert Einstein, volume 7* (pp. 306–320). Princeton, NJ: Princeton University Press.
- Einstein, A., & Pauli W. (1943). On the non-existence of regular stationary solutions of relativistic field equations. *Annals of Mathematics*, 44(2), 131–137.
- Finkelstein, D., & Rubinstein, J. (1968). Connection between spin, statistics, and kinks. *Journal of Mathematical Physics*, 9(11), 1762–1779.
- Fischer, A. E. (1970). The theory of superspace. In M. Carmeli, S.I. Fickler, L. Witten (Eds.), *Relativity*. Proceedings of the Relativity Conference in the Midwest, held June 2–6, 1969, at Cincinnati Ohio (pp. 303–357). New York: Plenum Press.

¹⁸Maxwell equations were thought of as a kind of effective theory that describes things on a coarse-grained scale, so that, e.g. the vortex knots could be approximated by point particles.

- Fischer, A. E. (1986). Resolving the singularities in the space of Riemannian geometries. *Journal of Mathematical Physics*, 27, 718–738.
- Friedman, J., & Mayer, S. (1982). Vacuum handles carrying angular momentum; electrovac handles carrying net charge. *Journal of Mathematical Physics*, 23(1), 109–115.
- Friedman, J., & Sorkin, R. (1980). Spin 1/2 from gravity. *Physical Review Letters*, 44, 1100–1103.
- Gannon, D. (1975). Singularities in nonsimply connected space-times. *Journal of Mathematical Physics*, 16(12), 2364–2367.
- Gibbons, G. W. (1968). The elliptic interpretation of black holes and quantum mechanics. *Nuclear Physics, B*, 98, 497–508.
- Gilbert, N. D. (1987). Presentations of the automorphisms group of a free product. *Proceedings of the London Mathematical Society*, 54, 115–140.
- Giulini, D. (1990). Interaction energies for three-dimensional wormholes. *Classical and Quantum Gravity*, 7(8), 1272–1290.
- Giulini, D. (1993). On the possibility of spinorial quantization in the Skyrme model. *Modern Physics Letters A*, 8(20), 1917–1924.
- Giulini, D. (1994). 3-manifolds for relativists. *International Journal of Theoretical Physics*, 33, 913–930.
- Giulini, D. (1995a). On the configuration-space topology in general relativity. *Helvetica Physica Acta*, 68, 86–111.
- Giulini, D. (1995b). Quantum mechanics on spaces with finite fundamental group. *Helvetica Physica Acta*, 68, 439–469.
- Giulini, D. (1995c). What is the geometry of superspace? *Physical Review D*, 51(10), 5630–5635.
- Giulini, D. (1997). The group of large diffeomorphisms in general relativity. *Banach Center Publications*, 39, 303–315.
- Giulini, D. (2007). Mapping-class groups of 3-manifolds in canonical quantum gravity. In B. Fauser, J. Tolksdorf, & E. Zeidler (Eds.), *Quantum gravity: Mathematical models and experimental bounds*. Basel: Birkhäuser Verlag. Online available at (arxiv.org/pdf/gr-qc/0606066).
- Giulini, D. (2008). Concepts of symmetry in the work of Wolfgang Pauli. In H. Atmanspacher & H. Primas (Eds.), *Recasting reality. Wolfgang Pauli's philosophical ideas and contemporary science* (pp. 33–82). Berlin: Springer. Online available at (arxiv.org/pdf/0802.4341v1).
- Giulini, D. (2009). The superspace of geometrodynamics. *General Relativity and Gravitation*, 41(4), 785–815.
- Gross, D., & Perry, M. (1983). Magnetic monopoles in Kaluza-Klein theories. *Nuclear Physics, B*, 115, 29–48.
- Hawking, S. W., & Ellis, G. F. R. (1973). *The large scale structure of spacetime*. Cambridge: Cambridge University Press.
- Hojman, S. A., Kuchař, K., & Teitelboim, C. (1973). New approach to general relativity. *Nature Physical Science*, 245, 97–98.
- Hojman, S. A., Kuchař, K., & Teitelboim, C. (1976). Geometrodynamics regained. *Annals of Physics*, 96, 88–135.
- Kazdan, J. L., & Warner, F. W. (1975). Scalar curvature and conformal deformation of Riemannian structure. *Journal of Differential Geometry*, 10(1), 113–134.
- Komar, A. (1959). Kovariant conservation laws in general relativity. *Physics Review*, 113(3), 934–936.
- Kruskal, M. D. (1960). Maximal extension of Schwarzschild metric. *Physical Review*, 119(5), 1743–1745.
- Kuchař, K. (1973). Geometrodynamics regained: A Lagrangian approach. *Journal of Mathematical Physics*, 15(6), 708–715.
- Lichnerowicz, A. (1955). *Théories Relativistes de la Gravitation et de l'Électromagnétisme*. Masson et Cie, Paris, 1955.
- McCarty, G. C., & Shultz, G. (1963). Homeotopy groups. *Transactions of the American Mathematical Society*, 106, 293–303.
- Misner, C. (1959). Wormhole initial conditions. *Physical Review*, 118(4), 1110–1111 (1959).
- Misner, C. (1963). The method of images in geometrostatics. *Annals of Physics*, 24, 102–117.

- Misner, C., Thorne, K. S., & Wheeler, J. A. (1973). *Gravitation*. New York: W.H. Freeman and Company.
- Misner, C., & Wheeler, J. A. (1957). Classical physics as geometry: Gravitation, electromagnetism, unquantized charge, and mass as properties of curved empty space. *Annals of Physics*, 2, 525–660.
- Müllner, D. (2008). *Orientation Reversal of Manifolds*. PhD thesis, Friedrich-Wilhelms-Universität Bonn, October (2008).
- Pesic, P. (Ed.), (2007). *Beyond geometry. Classic papers from Riemann to Einstein*. Mineola, NY: Dover Publications, Inc.
- Reidemeister, K. (1935). Homotopieringe und Linsenräume. *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg*, 11(1), 102–109.
- Riemann, B. (1869/1919). *Über die Hypothesen, welche der Geometrie zu Grunde liegen* (1869), (2nd ed.) (1919), Edited and annotated by H. Weyl. Berlin: Springer.
- Rindler, W. (1965). Elliptic Kruskal-Schwarzschild space. *Physical Review Letters*, 15(26), 1001–1002.
- Skyrme, T. H. R. (1971). Kinks and the Dirac equation. *Journal of Mathematical Physics*, 12(8), 1735–1743.
- Sorkin, R. (1977). On the relation between charge and topology. *Journal of Physics A: Mathematical and General*, 10(5), 717–725.
- Sorkin, R. (1983). Kaluza-Klein monopole. *Physical Review Letters*, 51(2), 87–90.
- Sorkin, R. (1988). A general relation between kink-exchange and kink-rotation. *Communications in Mathematical Physics*, 115, 421–434.
- Teitelboim, C. (1973). How commutators of constraints reflect the spacetime structure. *Annals of Physics*, 79(2), 542–557.
- Whitehead, J. H. C. (1941). On incidence matrices, nuclei and homotopy types. *Annals of Mathematics*, 42(5), 1197–1239.
- Witt, D. (1986). Symmetry groups of state vectors in canonical quantum gravity. *Journal of Mathematical Physics*, 27(2), 573–592.
- Woodhouse, N. (1991). *Geometric quantization* (2nd ed.). Oxford: Clarendon Press.

Chapter 13

The Surprising Resolution of the Poincaré Conjecture



Donal O'Shea

13.1 Introduction

In 2003, Grigory Perelman posted three papers (Perelman 2002, 2003a,b) on arXiv.org proving the Poincaré Conjecture, a century-old question that appears to be purely topological and that had become one of the most famous unsolved problems in mathematics. It asks whether the simplest topological property (simple-connectivity) characterizes the simplest closed three-manifold (the three-sphere). In what is surely one of history's greatest ironies, Perelman's proof draws heavily on analysis and geometry as well as methods developed in general relativity. It reunites strands of mathematics and physics that had grown out of staggering insights first set forth by Bernhard Riemann but which had diverged in the late nineteenth and early twentieth century. One strand gave rise to the absolute differential calculus and general relativity, a second led to the invention and understanding of homogeneous geometries, while a third strand eventually molted into algebraic, combinatorial and differential topology.

In Poincaré's hands, topology had developed into a pillar central to all of mathematics. In subsequent years, it would shed light on seemingly unrelated problems that involved equations from analysis and mathematical physics. No one, least of all Poincaré, would have ever imagined that techniques from analysis and mathematical physics to which topology had contributed so much, would one century later repay the favor by being used to solve the most famous purely topological problem of all time. And no one would have imagined that the object central to Perelman's proof, a hierarchy of Riemannian manifolds connected by

D. O'Shea (✉)

President's Office, New College of Florida, 5800 Bay Shore Road, Sarasota, FL 34243, USA

e-mail: doshea@ncf.edu

© Springer Science+Business Media, LLC, part of Springer Nature 2018

D. E. Rowe et al. (eds.), *Beyond Einstein*, Einstein Studies 14,

https://doi.org/10.1007/978-1-4939-7708-6_13

401

the Ricci flow, might provide a mathematical object useful for modeling space and space-time at different scales.

Perelman was awarded, but refused, the 2006 Fields medal for his work. After some initial controversy, his papers were accepted as correct, and the Poincaré Conjecture is now taken to be a theorem. A full account of his proof and a scrupulously detailed exposition of the parts of Perelman's work needed to establish it has been given by Morgan and Tian (2007). The author has written a history of the conjecture and the mathematics behind the proof (O'Shea 2007); see also (Szpiro 2007). The problem of proving or disproving Poincaré's conjecture was one of the seven "millennium problems" identified by the Clay Institute of Mathematics, each of which carries a cash prize of one million dollars for its solution. The Clay Institute announced in May, 2010 that Perelman would receive the first such award for his solution of the Poincaré Conjecture. He did not show up at the prize ceremony and subsequently declined the prize money. In what follows, we elaborate on the striking irony that a purely topological theorem, developed as part of a program realizing Poincaré's vision of topology as distinct from geometry and analysis, should require a proof that depends so heavily on the differential-geometric techniques that Riemann's vision foreshadowed and that play such a central role in general relativity. We begin by exploring why the Poincaré conjecture has exercised such a draw on the imagination of mathematicians.

13.2 Three-Dimensional Manifolds and the Poincaré Conjecture

Recall that a three-dimensional manifold is any topological space in which the neighborhood of each point is homeomorphic to a ball in Euclidean three-space \mathbf{R}^3 . So three-dimensional manifolds are sets that look locally like the empty space in our universe. They are of interest because they are mathematical models for the spatial structure of our universe: you can think of yourself living inside a three manifold. A particularly attractive class of three-dimensional manifolds consists of those that are *closed*: that is, compact (in particular, finite) and without boundary. There are uncountably many closed three-dimensional manifolds, but the simplest is the three-dimensional sphere, or three-sphere, which can be defined as any set homeomorphic to the set of points of fixed distance from the origin in Euclidean four-space. Topologists think of a three-sphere as the manifold that results from taking two disjoint three-dimensional solid balls and identifying corresponding points of the two-dimensional spheres that bound them. (This has the advantage of defining the three-sphere without requiring a four-dimensional space in which to embed it).

A topological space is *simply connected* if it is connected (hence, not the union of two disjoint open and closed subsets) and if every closed loop in the manifold can be continuously deformed within the space to its starting (and hence, end) point.

Both \mathbf{R}^3 and the three-sphere are easily seen to be simply connected. The *Poincaré conjecture*, now theorem, states that every closed, simply connected three manifold is homeomorphic to the three-sphere. At the time of its proof, the conjecture was one of the most famous unsolved problems in mathematics. It was one of the Clay Institute's seven millennium problems and one of only two that every mathematician whom the Institute canvassed had listed as one of the chief unsolved problems in mathematics – the other is the Riemann hypothesis,

What accounts for the problem's fame?

Any explanation must necessarily be speculative, but one that won't wash is the assertion that the conjecture, now theorem, owes its fame purely to Poincaré's stature. Poincaré was, to be sure, the most accomplished and influential mathematical scientist of his era. At the time of his death on July 17, 1912, he had been nominated for the Nobel prize 51 times.¹ However, Poincaré was extraordinarily prolific. In addition to a number of influential monographs, he published over 700 papers – more papers than the total number of pages published by Riemann. Many raised vital issues that begged resolution. A mere ten dealt with topology, and four of these (Poincaré 1892, 1899a, 1901b,c) were research announcements of the results contained in the other six. In a retrospective review of his own work (Poincaré 1901a) prepared at the request of Mittag-Leffler, Poincaré devoted only 4 of the 135 pages (pp. 100–103) to topology. And he does not mention the Poincaré conjecture.

In fact, the so-called Poincaré conjecture was not stated as a conjecture. Rather, Poincaré raises it as an easily overlooked question at the very end of his last topological paper (Poincaré 1904). Had Poincaré updated Hilbert's list of 23 problems, say in 1910, it is exceedingly doubtful that he would have listed his question as a major open problem. So Poincaré's fame cannot immediately account for his question's elevation to one of the best-known open problems in mathematics. Nor can the intuitive immediacy of three-manifolds or the simplicity of the conjecture's statement account for the hold that it later exercised. To fully understand why the conjecture became so famous, one needs to look at the larger influence of Poincaré's thought, and the emergence of topology into the mathematical mainstream in Europe and the United States.

13.3 Poincaré's Topological Papers

The notion that topology and geometry are distinct goes back to Bernhard Riemann's habilitation lecture (Riemann 1854). In making the distinction between topology and geometry, and in developing ways to bring the tools of analysis to bear on geometry, Riemann created the underpinnings for much mathematical progress

¹I owe the count to David Rowe, who extracted it from (Crawford et al. 1987), and who observes (private communication) that Darboux, who nominated Poincaré in seven different years, may have been behind a campaign to advance Poincaré's candidacy.

in the twentieth century. It was Poincaré, however, who fleshed out Riemann's topological ideas and welded them into a separate branch of mathematics.

As a student, Poincaré was interested in non-Euclidean geometry and studied differential equations. His mathematical education was rigorous, but insular. He knew of the work of Fuchs with whom he had corresponded, but he had no knowledge of Riemann's work. Indeed, relations between French and German mathematicians were never too wonderful, and more importantly their research traditions differed sharply. The rupture of relations between Germany and France following the Franco-Prussian war of 1870–71 had only widened the divide.

As a researcher at Caen, Poincaré noticed the relation between certain analytic objects and non-Euclidean geometry. More precisely, he saw the analogy between elliptic functions – which can be defined from a fundamental parallelogram in the Euclidean plane – and what he called Fuchsian functions, which are determined by their values on certain fundamental domains in the hyperbolic plane. His announcements drew the notice of Felix Klein, and subsequent correspondence with him led Poincaré to learn about and to absorb Riemann's approach to complex functions. The correspondence also touched off a rivalry between Poincaré and Klein, and the competition between the men ultimately led to the formulation of different types of uniformization theorems (Poincaré 1884) and (Klein 1883). The definitive version of these results, conjectured in the white heat of creativity resulting from their rivalry, and was finally proved in Poincaré (1907) and Koebe (1907). From this, the spectacular result that every closed surface (that is, two-dimensional manifold) carries a unique homogeneous geometry follows easily.²

Poincaré would duplicate his success a few years later. In 1887, King Oscar of Sweden had announced a large cash prize for a mathematical solution of the n -body problem (see Goroff 1993 and Barrow-Green 1997 for full accounts of the prize, the award, and a narrowly avoided, potentially career destroying, scandal.) Newton's law of gravity and his equations of motion provided a set of equations that, in principle, allowed one to calculate the position and velocity of each object in the solar system at any desired future time, once one knew the position, velocity, and mass of each object at some fixed time. But, in reality, one could not know the position, velocity, and mass of each object in the solar system exactly and, even if one did, the calculations would take too long on any conceivable computer (even today). There were, to be sure, techniques for making detailed perturbative calculations that served to predict where a given object would appear in the sky. But systematic theoretical understanding was out of reach. Even simplified models of the solar system with fewer planets or other objects were intractable. The behavior of Newton's equations for two heavenly bodies was well-understood, but the range of behavior in the next simplest case in which there were three bodies seemed to defy analysis. Was there a way to understand why, for example, most of the planets go around the sun roughly in the same plane? Or why the solar system does not fly apart and seems relatively stable? Or perhaps, things are not as they seem, and the

²For a careful account, see (Gray 2013, pp. 224–252).

solar system was doomed to self-destruct. (Laplace thought he had proved in his *Mécanique céleste* that this cannot happen.)

In order to attack this problem, Poincaré again turned to the topological concepts that had received their most powerful impetus in the then-thirty year old work of Riemann. These ideas, which had been further developed by Betti, Möbius, Peano and others, dealt with qualitative features of sets, such as their dimension and how they are connected. In his three-volume treatise *New Methods of Celestial Mechanics* (Poincaré 1892, 1893, 1899), which appeared between 1892 and 1899, Poincaré applied and reworked these concepts to address questions about the solar system. In so doing, he completely reframed the main questions and revolutionized the field. He transformed our notions of stability and discovered that even very simple systems could have behavior that was, in a precise sense, effectively unpredictable and chaotic. He discovered what later became known as chaos, and he exploited the power of topological concepts.

Poincaré's books and papers are marked by great breadth and a willingness to draw inspiration and methods from all fields of mathematics and physics. His topological papers, by contrast, are striking for their methodological purity. It was as if Poincaré decided to establish by example that topology could stand on its own as a discipline. Intentional or not, the strategy succeeded. Those six papers established topology as an independent discipline with its own methods and of interest in its own right. They would add immensely to then existing knowledge, and would establish new fields such as differential topology and algebraic topology that became central to twentieth century mathematics and the mathematics of today. Poincaré introduced algebraic concepts that allowed computation, and wielded topology into a discipline that would profoundly alter and enlarge the scope of all other areas of mathematics and their applications. He would later write that every mathematical or scientific problem he examined, no matter how remote or recondite, would lead him inexorably to topology.³ That experience would characterize much of the century's mathematics. Topology would go on to undergird many of the twentieth century's greatest advances.

Let us briefly review Poincaré's topological papers. His foundational paper on topology (Poincaré 1895) was announced in the Oct 31, 1892 volume of the *Comptes Rendus* (Poincaré 1892). The paper itself was the first to appear in the inaugural issue of the second series of the *Journal de l'École Polytechnique* in Paris. This journal dated back, with different names, to 1795 and had been published annually, albeit irregularly, since 1831. The second series marked its first centenary, and the placement of the paper indicates the importance that Poincaré and the journal editors accorded it.

The paper is extraordinary by any measure. It is 121 pages, long even by Poincaré's standards. In the introduction, Poincaré apologizes for its length, remarking that his attempts to shorten it led to obscurity. Although it is easy reading

³In a retrospective of his own work written in 1901, Poincaré writes: "As for me, all of the diverse paths on which I was successively engaged led me to analysis situs..." (Poincaré 1901b, p. 101).

for a mathematician today, as many of the concepts occur in the main graduate courses on topology and analysis, it must have been incredibly dense at the time. A short list of the topics covered will give some idea of its scope. Manifolds are first defined as the smooth zero loci of p smooth real valued functions of n real variables (usually $0 < p < n$). They are allowed to be bounded or not, and in the former case Poincaré gives a definition of the complete boundary. He also carefully defines homeomorphism. A second definition of manifold as a smoothly parameterized set in \mathbf{R}^n is given, and he carefully discusses the relation between the two definitions, showing that the second is more general than the first. He discusses analytic continuation in the case of analytic parameterizations and the necessity of restricting parameterizations to be one-to-one. He discusses orientation and oppositely oriented manifolds, and how orientation behaves with respect to defining conditions. He reviews homologies and Betti numbers. His discussion of the relation between the implicit and parametric definitions of manifolds allows him to define integrals of m -forms on m -dimensional manifolds. He sketches what we nowadays refer to as Stoke's theorem (the integral of an $(m - 1)$ -form over the complete boundary of an m -dimensional manifold is the integral of what is now called the exterior derivative of the form over the interior of the manifold), noting that this allows one to compute Betti numbers from periods of integrals of closed forms when the manifold is closed. He discusses what we now call non-orientable and orientable manifolds. A section on relative Betti numbers and what are now-called Kronecker indices follows together with a statement and proof of Poincaré duality. He states that he has never seen this theorem enounced, although he notes that it was certainly used by some individuals in applications. He then derives some easy consequences for the intersection form of the middle homology groups of even dimensional varieties in the case when the middle homology groups are even and odd dimensional. All this is in the first third of the article!

With these preliminaries out of the way, Poincaré turns to three-dimensional manifolds and geometric topology. He hangs on to the notion that three-dimensional manifolds are embedded in four space, but examines the creation of three dimensional manifolds by identifying suitable faces of chains of solid polyhedra. He examines the resulting cycles of vertices and edges, and notes that a condition for obtaining a manifold is that Euler's relation is satisfied for the star (which he defines) of the vertices. He carefully works out what happens when one identifies opposite faces of a cube in different ways (not failing to point out the analogy with the construction of the two dimensional torus and Klein bottle). He provides yet another construction of three-manifolds by looking at the quotient of discontinuous groups acting on three-space and shows how this relates to, and generalizes, his previous construction. In particular, he gives a class of examples of three-dimensional manifolds by looking at the quotient under a discrete group whose fundamental domain is the cube. Next, he introduces the fundamental group, relating it to deck transformations, and carefully sketching the differences from the first homology group. He discusses generators of the fundamental group, showing how to get them in the case where a three manifold is constructed by identifying opposite faces of a single polyhedron. He points out that these also give generators of the first

homology group, and works out the fundamental groups and Betti numbers of all the manifolds obtained by identifying opposite faces of the cube. He explicitly asks which invariants characterize a manifold up to homeomorphism, and uses the class of discontinuous groups he introduces to explicitly construct infinitely many three manifolds with the same Betti numbers.

After pointing out that not all Fuchsian groups⁴ are the fundamental groups of two-manifolds, he asks whether any group given by generators and relations can be the fundamental group of a closed n -dimensional manifold and, if so, how one can construct the manifold. He wonders whether two manifolds of the same dimension with the same fundamental group are necessarily homeomorphic, a question that seems astonishing naïve given the incredible range and sophistication of the paper. He devotes the last third of the article to some other ways of creating manifolds, working out some complicated examples. He generalizes the Euler formula to what is now known as the Euler-Poincaré formula for polyhedra and manifolds of any dimension, giving two proofs.

In summary, in this foundational paper (Poincaré 1895), Poincaré lays out the basics of what would become differential topology, algebraic topology, and combinatorial topology. He covers a vast amount of territory, and even today the contrast between how far he gets and what he does not know is breathtaking. Despite his deep penetration into the subject, he does not, for example, have access to the simple examples that students of today construct using products (or connected sums) of manifolds.

The five papers, which he calls compléments, that follow contain even more material than the foundational paper and immeasurably deepen the examples and techniques. The second (Poincaré 1899b) clarifies the definition of Betti numbers, and responds to Heegaard's criticism of his proof of Poincaré duality. Poincaré introduces triangulations and attempts to show that every manifold can be triangulated. The third (Poincaré 1900) introduces torsion coefficients and shows that Poincaré duality holds for them as well. He states as a theorem (without proof, as the claim is false) that any three-dimensional manifold with the homology of the sphere is homeomorphic to the sphere. In the fourth paper (third complément: Poincaré 1902a), he studies the topology of algebraic surfaces of the form $z^2 = F(x, y)$ where x, y, z are complex and the curve $F(x, y) = 0$ is nonsingular or possesses at most ordinary double points. The fifth paper (Poincaré 1902b) examines general algebraic surfaces $F(x, y, z) = 0$, elaborating on work of Picard. A good chunk of Picard-Lefschetz theory can be found here.

Poincaré's sixth (Poincaré 1904) and final topological paper, which appears in 1904, introduces what we now call Morse theory. Poincaré investigates the role of surface diffeomorphisms in gluing handlebodies and creating three-manifolds. By attaching two genus two handlebodies, Poincaré constructs a beautiful manifold, often called the Poincaré dodecahedral space (although the connection with the dodecahedron was not established until years later by Kneser), with finite funda-

⁴Recall that Fuchsian groups go back to Poincaré's work in function theory from the early 1880s.

mental group that is a homology sphere.⁵ He provides thereby a counter-example to the “theorem” he had announced in his third paper. Immediately after presenting this example, on the last page of the paper, Poincaré asks whether a closed simply connected three-dimensional manifold is homeomorphic to the three-sphere. This, of course, is the Poincaré conjecture. He ends abruptly with the sentence: “But this question would take us too far afield.”

In his first foundational paper, Poincaré had found it necessary to defend the use of higher dimensions. Throughout the six papers, he devotes most of his attention to three and four dimensional manifolds, but is always mindful of higher dimensions. The increase in sophistication between the first paper and the sixth is enormous. By the last paper the field of topology is fully launched. It would be many years until even specialized textbooks caught up with the material Poincaré covered, and even today there is no textbook with the same scope as Poincaré’s six papers.

It bears repeating that unlike his analytic papers, which mix geometry, analysis, and physics, Poincaré does not mix categories in the topological papers. Even though he ranges over all areas of manifold topology, he is always careful to confine himself to topological techniques. It is almost as if he decided to underscore Riemann’s distinction between topology and geometry, and consciously kept them distinct. Whatever the reason, the purely topological treatment of topology allowed the field to develop in its own right and established it as a separate branch of mathematics. So, the cumulative effect of Poincaré’s papers was to establish topology as a discipline in its own right, and its central status in mathematics grew as the twentieth century progressed. Topology also took hold in the United States, and was central to the emergence of the American mathematical research community, particularly in America’s great land grant universities in the Midwest.

From this point of view, the conjectural, purely topological characterization of the three-sphere was irresistible. It was arguably the simplest question that one could ask about a sphere (although it took some time to ask). It was easy to state and three manifolds were easy to construct. Moreover, topology’s success lulled almost everyone into believing that the Poincaré conjecture must have a purely topological proof. And perhaps some day, someone will find one. Certainly, Poincaré’s last paper provided enough promising techniques to have lured some mathematicians into devoting their whole careers to working on the conjecture. Little by little, they got pulled in and then years would go by.

As time went on, it became clearer that the unresolved conjecture was quite treacherous and very subtle. Indeed, in some ways, Poincaré was its first victim. At the end of his third major topological paper (Poincaré 1900, p. 308), he formulated a related assertion (that any closed manifold with the homology of a sphere – that is, vanishing first and second homology groups – is homeomorphic to a sphere) as a theorem. He subsequently discovered that this “theorem” was false, and a good portion of his last topological paper was devoted to constructing a lovely counter-

⁵For a conjectural account of how Poincaré might have found his counterexample, see Jeremy Gray’s account in Gray (2013).

example. Poincaré's error, which the last part of his last paper corrects, certainly contributed to the final question from that paper becoming one of the most avidly sought after results in mathematics.

Far less clear in 1904 was how difficult answering this question would turn out to be. The Poincaré conjecture would prove to be one of the most vexing open problems in mathematics, and it resisted all attempts to prove or disprove it for a century. Proposed proofs were found to contain irremediable gaps, and purported counter-examples fell apart under closer intense scrutiny. The failures increased the conjecture's fame, and the unresolved conjecture became an alluringly beautiful siren song that lay at the heart of a troubled century's most splendid achievements. To add insult to injury, its analogue in all dimensions greater than four of was proved in 1960 (Smale 1961), and in dimension four in (Freedman 1982). In both cases, these breakthroughs caused a sensation, drawing the highest awards in mathematics, and spurring new progress in topology. Nevertheless, the circumstances that allowed the proof to go through in dimensions greater than or equal to five could not be extended to lower dimensions, and the four-dimensional methods were specific to that case. So the lack of a resolution in dimension three, the case with the most immediate implications for the global topology of our universe, continued to be a source of embarrassment for topologists. There were even serious doubts about whether the three dimensional case was true. Yet today these skeptics have vanished from the scene: the Poincaré conjecture, now a theorem thanks to Perelman, provides a simple, purely topological characterization of the three-sphere.

13.4 Perelman's Proof

Perelman's proof of the Poincaré conjecture is a spectacular mix of differential geometry and analysis combined with pure topology that builds on two major flights of ideas of the late twentieth century. The first is Bill Thurston's discovery that there are only eight essentially different homogeneous geometries that a compact three-manifold can possess. Homogeneous means that all points of the manifold look the same in the sense that a ball of sufficiently small fixed radius about one point is isometric to the balls about any other point of the same radius. This is the analogue of the great nineteenth century discovery that there are three distinct types of homogeneous geometries in dimension two, a discovery that definitively ended the two-thousand year old controversy about Euclid's parallel postulate. The geometries in two dimensions can be characterized in (at least) two essentially different ways. One can proceed axiomatically, according as to whether there are no, exactly one, or infinitely many parallel lines through a given point not on a line (in which case every triangle has, respectively, angle sum greater than, equal to, or less than π , in which case the geometry is called spherical, Euclidean, or hyperbolic, respectively), or one may take the differential geometric approach, in which case one studies surfaces of constant curvature and differentiates between them according to whether the curvature of the surface is greater than, equal to, or less than zero.

Simple examples show that not every three-dimensional manifold can carry a homogeneous geometry, but Thurston nevertheless conjectured (and proved for a large class of manifolds) that every closed three-dimensional manifold could be cut into simpler three manifolds in a natural manner, so that each piece carries one of the eight possible homogeneous geometries. Were this conjectural result correct, then it would imply the Poincaré conjecture, but unfortunately a proof of Thurston's conjecture seemed far out of reach. It did, however, give some reason to believe that the Poincaré conjecture might be true. Thurston's conjectured result was the analogue in dimension three of the result that every closed two-dimensional surface carries one (and only one) of the three types of homogeneous geometries. This result, one of the loveliest of nineteenth-century mathematics, is inextricably connected with the discoveries of automorphic and Fuchsian functions that Poincaré made early in his career and that made him famous.

The second main flight of ideas on which Perelman built belongs to analysis, where he drew on the work of Richard Hamilton and his coworkers dealing with Ricci flows. Following an idea that goes back to the late James Eells, one uses the fact that any differentiable manifold can be given a Riemannian metric (and any three-manifold manifold is necessarily a differentiable manifold – see below). Given such a metric, one can define lengths of tangent vectors and angles between them, hence distance and angles on the manifold. In fact, one also can introduce all the paraphernalia of differential geometry, including the Riemann tensor, sectional curvatures, Ricci curvature, and scalar curvature.

Our account so far glosses over much history and many technicalities. We pause to provide more detailed references to both. Historically, it took many decades to clarify the notion of a manifold (see Scholz 1980, 1999), and even longer to clarify the relations between the different rigorous definitions of manifolds. The manifolds we have defined are topological manifolds. A differentiable manifold is a (topological) manifold on which one can do calculus. Another class of manifolds, the so-called piecewise-linear manifolds, arises from gluing together polyhedra. All three types of manifolds – topological, piecewise-linear, and differentiable – can be found in Poincaré's first foundational paper (Poincaré 1895). It is not the case that every manifold can be given the structure of a differentiable manifold (Kervaire 1960), but it is the case in dimension three. The latter follows from the fact that every three-dimensional manifold has an essentially unique piecewise-linear (Moise 1952), thus differentiable (Munkres 1960), structure. That a differentiable manifold has a Riemannian metric is established by a standard “partition of unity” argument. See, for example, (Kosinski 2007) which contains an excellent account of the fundamentals of differentiable topology and many historical notes. Consult (Milnor 2003) for an account of some of the approaches to proving the Poincaré conjecture.

To return to our account, let us then imagine that one has a three-manifold with a Riemannian metric. Hamilton considered how one might improve the metric. In particular, he considered a one parameter family $g(t)$ of Riemannian metrics, with $g(0)$ being the metric with which one starts, and sought to write out an equation of the form $g'(t) = F(g(t))$ where F is some functional of the metric that smooths

it appropriately. The fact that F depends only on g reflects the desideratum that the smoothing mechanism depend only on the metric and not on the coordinates utilized. For each t the Riemannian metric on the left side of the equation is a contravariant symmetric two-tensor, that is a real-valued function that, at each point of the manifold, assigns a number to any two tangent vectors at that point. Thus, the right hand side must be the same sort of object. If we stipulate that it not depend on derivatives of the metric higher than second order, then F is forced to be a linear combination of the metric tensor itself and the Ricci curvature. (There are no other inequivalent contravariant symmetric two-tensors that depend only on the metric that are second order or less – see, for example, Anderson (2004).) The same considerations apply in dimension four, although this was not realized at the time that general relativity was being formulated. This accounts for why the Ricci tensor appears in the Einstein equations and plays such a large role in general relativity. Hamilton focused on the case $g'(t) = -2Ric(g(t))$, where the constant does not matter as it can be rescaled.

Hamilton showed that the Ricci flow behaved particularly well with respect to Thurston's geometries. In particular, if one applies the Ricci flow to a manifold having one of the eight different homogenous geometries, then the geometry stays the same except that distances rescale by a constant factor. He showed that for Ricci flows on two-dimensional manifolds, the metric tends to one with the homogeneous geometry supported by the manifold, thereby reproving Poincaré's result that every closed two-dimensional manifold has a unique geometry. He further showed that by starting with a manifold in which the Ricci curvature was positive and bounded away from zero, then under the Ricci flow the metric tends to one giving a homogeneous spherical geometry. Hamilton hoped that a Ricci flow could be applied to an arbitrary three-dimensional manifold and that this might allow one to prove Thurston's conjecture. However, the technicalities were daunting. Even if one started with a manifold that admitted a homogeneous metric, the metrics that evolved in the course of a Ricci flow could develop singularities. On more complicated manifolds that were composed of manifolds with different geometries, the metric converged at different speeds in different parts of the manifold, leading to seemingly intractable singularities. Hamilton had made some progress defining a so-called "Ricci flow with surgery" in which one considers a Ricci flow on a manifold, cuts out regions of the manifold as they develop singularities to obtain a new manifold, and successively restarts the flow on the new manifold. By the mid 1990's, however, a consensus emerged that the Hamilton program was a powerful tool, but that it would not (and could not be made to work to) establish Thurston's conjecture.

Perelman's work changed all this. He introduced a number of new mathematical concepts that enabled him not only to explore and tame the singularities that developed in the Ricci flow but also to define the Ricci flow with surgery. Amazingly, the Ricci flow with surgery appears to be the perfect tool for endowing regions of a manifold with nice geometries and carving it up into topologically simpler components. If one applies the Ricci flow to a three-dimensional manifold that is already simply connected, Perelman showed that his arguments simplified and

provided a direct proof of the Poincaré conjecture. Moreover, many of Perelman's results on the Ricci flow apply quite generally, raising the promise of finding new results in higher dimensions.

Too little time has elapsed from the appearance of Perelman's posts to assess the influence of his work, much less evaluate the social context and initial controversy surrounding it. Nonetheless, it seems clear that Perelman's work will stimulate the efflorescence of three-manifold topology and geometry, and that Ricci flow methods will revolutionize areas of topology, general relativity and physics. It also seems likely that future historians of mathematics and science will find much to ponder and explain in the rapidity with which consensus has built around the validity of Perelman's arguments. Here is why.

The Poincaré conjecture is a purely topological statement, and for reasons outlined in the preceding section, many mathematicians had come to expect that a proof would be purely topological. The problem was also known to be ferociously tricky. Many attempts at proving or disproving the conjecture turned out to have subtle errors, and much of the folklore surrounding the failed attempts had been shared widely. Like most complex communities, the mathematical community is made up of many different networks of researchers and students. However, work on the conjecture in the sub-communities most likely to make progress had stalled, and the groups did not entirely trust one another. Despite the successes of Thurston's program, many low-dimensional topologists still resented the invasion of geometric techniques into their field. Among the three-dimensional topologists and geometers who welcomed the geometric techniques, there were many who resented the incursion of differential geometers wielding highly technical, seemingly impenetrable, and very delicate, hence fallible, tools drawn from partial differential equations. And, indeed, some of these differential geometers seemed like analysts whose sympathies and instincts were neither geometric nor topological.

In such a context, opportunities for misunderstanding and mischief abound. Yet Perelman's complicated, technically brilliant results were broadly accepted as correct within three years of his first post. By any accounting, that is astonishingly fast. The wonder is not that there was controversy, or that it spilled over into the mainstream press, but that there was not more. A heartfelt priority dispute with very capable mathematicians on either side could have escalated, instead of being resolved quickly and decisively.

Part of the explanation for the rapid acceptance of the proof involves the almost-immediate emergence of several groups devoted to checking, and enlarging upon, the details of Perelman's arguments. These groups were greatly aided by the immediacy of electronic communication. A web site organized by John Lott and Bruce Kleiner, formerly hosted at the University of Michigan and now hosted at UC Berkeley, drew contributions from the largely, but not exclusively, American community of topologists, geometers and differential geometers, many influenced by Thurston. Another group organized by Gerard Besson, who hosted a web site at the University of Grenoble that drew contributions from many members of the European community. A group of very analytically inclined differential geometers and specialists on the Ricci flow formed around Shing-Tung Yau at Harvard.

Needless to say, electronic communities and communication are not the whole story, and there is much to be said about the differing approaches of the groups and their internal dynamics. The Clay Institute, which had sponsored the millennium prizes, also played a key role in the dissemination of Perelman's results. The rapid settlement of priority claims suggests discreet and skillful leadership within the mathematical community. Perhaps even more importantly, some well known mathematicians who were not specialists in Ricci flow technology took time from their research to master and write up Perelman's work.

Although there can be no doubt that Perelman was interested in proving the Poincaré conjecture (and beyond it, the Thurston geometrization conjecture), this was not his only goal. Perelman's remarks at the outset of his first preprint (Perelman 2002), and the possible insight they give into his motivation, have gone curiously unremarked. In particular, Perelman raises the possibility of his work allowing one to replace the notion of a manifold as the underlying mathematical model of the universe with a hierarchy of manifolds at different scales. These manifolds might differ topologically from one another, but they would be connected by means of the Ricci flow. By way of analogy, the surface of the earth, viewed from far away, has the topology of a sphere. But as one approaches closer and closer, other features appear, such as natural (and man-made) bridges that complicate the topology with small, multiply connected handles. So, too, as one looks at the universe with finer and finer scales, it seems likely that the topology (and certainly the geometry) becomes more complex – think of localized singularities caused by black holes. Very speculatively, Perelman suggests that one could attempt to model the passage to larger scales as a Ricci flow with scale as a parameter. One could imagine doing this with space-like Riemannian three-dimensional manifolds or with four-dimensional Lorentzian manifolds modeled on space-time.

13.5 Conclusion

Poincaré's famous work on automorphic functions had started by uniting geometry and topology, and his early fame rested on those achievements. From that point of view, it is not so surprising that the Poincaré conjecture might involve geometry and, in retrospect, the real surprise in Perelman's proof is that it was such a surprise. Poincaré's success in establishing topology as a freestanding field, distinct from geometry, had lulled many mathematicians into expecting that the resolution of the Poincaré conjecture would be purely topological. Ironically, what undergirds Perelman's proof – the fact that three-dimensional manifolds have pieces that have unique geometries – is the three-dimensional analogue of Poincaré's discovery that a closed two-dimensional manifold has a unique geometry.

What is most satisfying about Perelman's work is the way in which it unites various areas of mathematics that had very productively gone their separate ways in the aftermath of Riemann's and Poincaré's insights. The late nineteenth and early twentieth century saw the separate development of the strands that gave rise to the

absolute differential calculus and general relativity, the strand that gave rise to the invention and understanding of homogeneous geometries, and the strand that gave rise to algebraic, combinatorial and differential topology. All trace back directly to Riemann. In Poincaré's hands, topology became central to mathematics. In subsequent years, it would shed light on seemingly unrelated problems that involved equations from analysis and mathematical physics. It underlies the exotic string theories at the edge of modern cosmology and is the science behind the construction of internet search algorithms, such as those used by Google. No one, least of all Poincaré, would have ever imagined that techniques from analysis and mathematical physics to which topology had contributed so much, would one century later repay the favor by being used to solve the most famous purely topological problem of all time. And no one would have predicted that the resulting gadget, a hierarchy of Riemannian manifolds connected by the Ricci flow, might provide a mathematical object useful for modeling space and space-time at different scales. What goes around, comes around.

References

- Anderson, M. (2004). Geometrization of 3-manifolds via the Ricci flow. *Notices of the American Mathematical Society*, 51, 184–193.
- Barrow-Green, J. (1997). *Poincaré and the three-body problem*. Providence: American Mathematical Society.
- Crawford, E., Heilbron, J. L., & Ullrich, R. (1987). *The noble population 1901–1937: A census of the nominators and nominees for the prizes in physics and chemistry*. Berkeley: Office for the History of Science and Technology, University of California, Berkeley.
- Freedman, M. (1982). The topology of four-manifolds. *Journal of Differential Geometry*, 17, 357–454.
- Goroff, D. (1993). Henri Poincaré and the birth of chaos theory: An introduction to the English translation of *Les Méthodes nouvelles de la mécanique céleste*. In D. Goroff (Ed.), *New methods of celestial mechanics* (1st ed., pp. 11–1107). College Park: American Institute of Physics.
- Gray, J. (2013). *Henri Poincaré: A scientific biography*. Princeton: Princeton University Press.
- Kervaire, M. (1960). A manifold which does not admit any differentiable structure. *Commentarii Mathematici Helvetici*, 34, 257–270.
- Klein, F. (1983). Neue Beiträge zur Reimannschen Funktiontheorie. *Mathematische Annalen*, 21, 141–218.
- Koebe, P. (1907). *Zur Uniformisierung der beliebiger analytischer Kurven*. *Nachr. K Ges Wissenschaft. Göttinger Math Phys Kl.*, 191–210. (Also II (1907) 177–198; III (1908) 337–358; IV (1909) 324–361.)
- Kosinski, A. (2007). *Differential manifolds*. Mineola: Dover. (Second edition of 1993 original published by Boston: Academic Press.)
- Milnor, J. W. (2003). Towards the Poincaré conjecture and the classification of 3-manifolds. *Notices of the American Mathematical Society*, 50, 1226–1233.
- Moise, E. (1952). Affine Structures in 3-manifolds. V. The triangulation theorem and Hauptvermutung. *Annals of Mathematics*, 56, 96–114.
- Morgan, J., & Tian, G. (2007). *Ricci flow and the Poincaré conjecture*. Providence: American Mathematical Society.

- Munkres, J. (1960). Obstructions to the smoothing of piecewise-differentiable homeomorphisms. *Annals of Mathematics*, 72, 521–554.
- O’Shea, D. (2007). *The Poincaré conjecture: In search of the shape of the universe*. New York: Walker Books.
- Perelman, G. (2002). The entropy formula for the Ricci flow and its geometric applications. arXiv: math/0211159v1 [math.DG].
- Perelman, G. (2003a). Ricci flow with surgery on three-manifolds. arXiv: math/0303109v1 [math.DG].
- Perelman, G. (2003b). Finite extinction time for the solutions to the Ricci flow on certain three-manifolds. arXiv: math/0307245v1 [math.DG].
- Poincaré, H. (1884). Sur les groupes des equations linéaires. *Acta Mathematica*, 4, 201–311.
- Poincaré, H. (1892). Sur l’analysis situs. *Comptes rendus de l’Académie des Sciences*, 115, 663–666.
- Poincaré, H. (1892, 1893, 1899). *Les Méthodes Nouvelles de la Mécanique Céleste* (Vols. I, II, III). Paris: Gauthier-Villars.
- Poincaré, H. (1895). Analysis situs. *Journal de l’École Polytechnique*, 1, 1–121.
- Poincaré, H. (1899a). Sur les nombres de Betti. *Comptes rendus de l’Académie des Sciences*, 128, 629–630.
- Poincaré, H. (1899b). Complément à l’analysis situs. *Rendiconti del Circolo Matematico di Palermo*, 13, 285–343.
- Poincaré, H. (1900). Second complément à l’analysis situs. *Proceedings of the London Mathematical Society*, 32, 277–308.
- Poincaré, H. (1901a). Analyse des travaux scientifiques de Henri Poincaré faite par lui-même. *Acta Mathematica*, 38(1921), 1–135.
- Poincaré, H. (1901b). Sur l’analysis situs. *Comptes rendus de l’Académie des Sciences*, 133, 707–709.
- Poincaré, H. (1901c). Sur la connexion des surfaces algébriques. *Comptes rendus de l’Académie des Sciences*, 133, 969–973.
- Poincaré, H. (1902a). Sur certaines surfaces algébriques : troisième complément à l’analysis situs. *Bulletin de la Société Mathématique de France*, 30, 49–70.
- Poincaré, H. (1902b). Sur les cycles des surfaces algébriques : quatrième complément à l’analysis situs. *Journal de Mathématiques*, 8, 169–214.
- Poincaré, H. (1904). Cinquième complément à l’analysis situs. *Rendiconti del Circolo Matematico di Palermo*, 18, 45–110.
- Poincaré, H. (1907). Sur l’uniformisation des fonctions analytiques. *Acta Mathematica*, 31, 1–63.
- Riemann, B. (1854). *Über die Hypothesen, welche der Geometrie zu Grunde liegen*. Habilitationvortrag. Göttingen. Göttinger Abhandlungen 13 (1867).
- Scholz, E. (1980). *Geschichte des Mannigfaltigkeitsbegriffs von Riemann bis Poincaré*. Boston: Birkhäuser.
- Scholz, E. (1999). The concept of a manifold, 1850–1950. In I. James (Ed.), *History of topology* (pp. 25–64). Amsterdam: Elsevier.
- Smale, S. (1961). The generalised Poincaré conjecture in dimensions greater than four. *Annals of Mathematics*, 74, 391–406.
- Szpiro, G. (2007). *Poincaré’s prize: The hundred-year quest to solve one of math’s greatest puzzles*. New York: Dutton.

Chapter 14

Unified Field Theory up to the 1960s: Its Development and Some Interactions Among Research Groups



Hubert Goenner

14.1 Introduction

Why should we look back into the history of (classical) unified field theory (UFT)? Research in this field as carried on during the three decades up to the 1960s has not succeeded to describe all fundamental interactions. Moreover, today we think that unification of the fundamental fields cannot be realized without using local quantum field theory. Whether this will happen in a more general geometrical framework than Minkowski space, if in any, is still an open question. Nevertheless, there are at least three motives for the study of the history of UFT:

1. The model case of the geometrization of gravitation by its identification with objects of Riemannian geometry was so successful that an extension in *some* direction seems not altogether absurd. From hindsight we see that the attempted joinder of gravitation and electromagnetism, as UFT wanted it, just could not work. The unification of the electroweak and strong interactions has fared much better.
2. A second interest might be to look into the conceptual and methodical framework of this particular scientific community. Why did the theoretical physicists involved do what they did? Were they:
 - trying to explain nature?
 - rating mathematics higher than physics? The outcome makes us recognize that mere trust in powerful mathematical methods without an intense link to experimental/observational data does not guarantee progress for physics.

H. Goenner (✉)

University of Göttingen, Institute for Theoretical Physics, Friedrich-Hund-Platz 1, D-37077 Göttingen, Germany

e-mail: Hubert.Goenner@Physik.Uni-Goettingen.DE

© Springer Science+Business Media, LLC, part of Springer Nature 2018

D. E. Rowe et al. (eds.), *Beyond Einstein*, Einstein Studies 14,

https://doi.org/10.1007/978-1-4939-7708-6_14

417

- trying to avoid the quantum realm?
- just following an authority like Albert Einstein?

UFT is a field rich in hypotheses if not in sheer *ideology* influencing its course. Think at the assumption that unification and geometrization be inseparable, an idea followed by the mainstream of researchers in this field. Still worse was the clinging to the even less secured hypothesis that geometrization must be done within *mixed* geometry, i.e., metric-affine geometry with an asymmetrical metric!

Examples for alternative approaches with an only partial geometrization would be Riemannian geometry with extra fields without direct geometrical meaning (classical Dirac field, scalar fields, or matter fields in general). In UFT this alternative course was excluded in the beginning, because the material sources of fields were supposed to also become part of the geometry. In the course of the development, matter variables again were added by some researchers.

3. How research in UFT was organized? Mostly one-man-endeavours or collaboration within groups? From the point of view of the sociology of science, a study of the interaction of people in such a relatively small-sized subcommunity might be interesting:
 - Who were the various research groups?
 - How did they communicate among each other and with the physics community as a whole, i.e., by traveling to each other or to conferences, by publications in scientific journals?

In this article, I will try to give partial and preliminary answers to some of the questions asked. At first, in Section 14.2 an outline of the field will be given with main actors Albert Einstein (1879–1955), Erwin Schrödinger (1887–1961), Marie-Antoinette Tonnelat (1912–1980), and Vaclav Hlavatý (1894–1969).¹ In Section 14.3, my focus will be on the original choice of the field equations and the ambiguities inherent in UFT. Attempts for their solution will be discussed in Section 14.3.1. How the field equations were changed in the course of time is seen in Section 14.3.2.

There is a close relationship among:

- the geometry taken as a framework;
- the freedom in the choice of field equations within this geometry;
- the linking of geometrical objects to physical observables;
- the gaining of results usable for the description of physical systems. In the course of time the need for a change in the field equations occurred enforced by negative physical results.

¹The Italian groups around Bruno Finzi (1899–1974) and Maria Pastori (1895–1975) will be taken into account in Section 14.4.2.

In the last Section 14.4, a very brief presentation will be given of some of the many scientists of the worldwide community involved and their contributions to theory building. This paper draws on the two papers (Goenner 2004) and (Goenner 2014) published in *Living Reviews in Relativity*, which cover the period 1920 to 1965.²

14.2 The Geometrization of Physics

When physical structures are to be modeled with the help of differential manifolds like space-time or fibre bundles, we encounter basic geometrical concepts, like “metric”, (linear or affine) “connection” and symmetries connected to a group structure. The metric with its components g_{ik} ($i, k = 0, 1, 2, 3$) is defined by a quadratic form $g(X, X) = g_{ik}X^iX^k$ and represents a *measure of distance* between two infinitesimally neighbouring events with coordinates x^i and $x^i + dx^i$: $ds^2 = g_{ik}dx^i dx^k$; it also governs the norm $|Y| := \sqrt{|g_{ij}Y^iY^j|}$ of a 4-vector Y , and an “angle” $\cos(\angle(X, Y)) := \frac{g_{ij}X^iY^j}{|X||Y|}$ between two such vectors X, Y .

A Lorentz metric (i.e., with signature ± 2) determines the lightcones as well, or the *causal structure* of space-time, up to a conformal factor:

$$g(X, X) = 0. \tag{14.1}$$

So far we have tacitly assumed the *symmetry* of the metric: $g(X, Y) = g(Y, X)$. In the following, we also will have to cope with an *asymmetric* metric. Its components may be decomposed into symmetrical and skew-symmetrical parts:

$$g_{ik} = h_{ik} + k_{ik} \tag{14.2}$$

with $h_{ik} = g_{(ik)} =: 1/2(g_{ik} + g_{ki})$, $k_{ik} = g_{[ik]} =: 1/2(g_{ik} - g_{ki})$.³

From a geometrical perspective, the mere addition of a quadratic form $g(X, X)$ and a 2-form $k = k_{ij}dx^i \wedge dx^j$ in building a new field variable is less than convincing. In a standard interpretation, a term describing gravitational *potentials* is added to a term standing for electromagnetic *fields*. Moreover, both h_{ik} and k_{ij} , each constitute an *irreducible* representation with regard to the permutation group. Of course, the formation of the *inverse* of the metric via $g^{il}g_{kl} = \delta_k^i = g^{li}g_{lk}$ intertwines the symmetric and skew-symmetric parts.⁴ In fact, A. Lichnerowicz

²This 2nd part will, hopefully, be ready during 2010.

³The skew-symmetrical part appears in the “angle” and leads to an unwanted property: The angle between X and Y is *different* from the angle between Y and X : $\cos(\angle(X, Y)) - \cos(\angle(Y, X)) = 2 \frac{k_{ij}X^iY^j}{|X||Y|} \neq 0$.

⁴Note that the symmetric and skew-symmetric parts have their own inverses.

(1915–1998) favoured the symmetrical part l^{ik} of the inverse asymmetric metric $g^{ik} = l^{ik} + m^{ik}$, or its inverse l_{ij} ($l_{is}l^{ks} = 0$) to stand for the metric (Lichnerowicz 1955a, 288; Lichnerowicz 1955b). This resulted from his study of the Cauchy initial value problem and his investigation of whether the Einstein-Straus-system of field equations was free of inconsistencies (Lichnerowicz 1953, 1954). A geometry with symmetric metric and an affine connection is called metric-affine geometry; if the metric is asymmetric I call it *mixed* geometry. It is mixed geometry which underlies much of the research in UFT during the decades looked at. If the connection is the only dynamical variable by help of which a metric then can be defined, I will speak of *purely affine* geometry.

The second fundamental object, the affine *connection* L is a prescription of how to transport vectors *parallelly* from one arbitrary neighbouring event p to another q on the manifold or from one fibre to another one in a fibre bundle.⁵ For the components Y^k of a 4-vector, the prescription is the following:

$$\delta Y^k =: Y^k(q) - Y^k(p) = -dx^l L_{lm}{}^k Y^m(p) + O(dx^2). \quad (14.3)$$

$L_{lm}{}^k$ are the components of the affine *connection* L . The concepts “metric” and “connection” are independent of each other.

By help of the connection, *tensorial* derivatives, named *covariant* derivatives can be introduced:

$$Y^i{}_{||k} =: \frac{\partial Y^i}{\partial x^k} + L_{kl}{}^i Y^l \quad (14.4)$$

vid.

$$\omega_i{}_{||k} =: \frac{\partial \omega_i}{\partial x^k} - L_{ki}{}^l \omega_l. \quad (14.5)$$

Parallel transport of a vector Y along a curve $x^k(u)$ with tangent vector X is defined by:

$$\nabla_X Y = 0 \Leftrightarrow Y^i{}_{||k} \frac{dx^k}{du} = 0. \quad (14.6)$$

If the connection is decomposed into a symmetric and a skew-symmetric part: $L_{lm}{}^k = L_{(lm)}{}^k + L_{[lm]}{}^k$, then the symmetric part again is a connection while the skew-symmetric part is a tensor called *torsion* tensor. Above, I used the plural “derivatives” because with $L_{ij}{}^k$ also $L_{(ij)}{}^k$ is a connection due to the decomposition

⁵For an arbitrary curved surface there is no a priori rule when to call vectors in different points as being parallel.

$$L'_{ij}{}^k := L_{(ij)}{}^k + T_{ij}{}^k . \tag{14.7}$$

Here, $T_{ij}{}^k$ is an arbitrary tensor field. This remark is not as trivial as it appears if the connection is asymmetrical. By taking $T_{ij}{}^k = \pm S_{ij}{}^k$ with the torsion tensor S , we may interpret $L_{ij}{}^k = L_{(ij)}{}^k + S_{ij}{}^k$ and $L_{ji}{}^k = L_{(ij)}{}^k - S_{ij}{}^k$ as two connections appearing in Einstein’s famous condition for the (asymmetric) metric:

$$g_{ik||l} :=: g_{ik,l} - g_{rk}L_{il}{}^r - g_{ir}L_{lk}{}^r = 0 . \tag{14.8}$$

Equation (14.8) is replacing the condition well known to us from Riemannian geometry:

$$g_{ik||l} :=: g_{ik,l} - g_{rk}\{^r_{il}\} - g_{ir}\{^r_{kl}\} = 0 . \tag{14.9}$$

We call Equation (14.8) the “compatibility” equation, although it does *not* imply the beneficial consequences metric compatibility (14.9) has in Riemannian geometry.⁶ If for a symmetric metric the tensor $g_{ij||k} := Q_{ijk}$, named “non-metricity tensor” Q , is given a certain value, then metric and connection are related. In Riemannian geometry $Q = 0$; the connection is called *metric-compatible*. In this case, norms and angles are both conserved under parallel transport.

Already in 1918, the mathematician *Hermann Weyl* suggested a UFT with $Q \neq 0$ (Weyl 1919). In his theory an arbitrary re-calibration or gauging (“Umeichung”) of clocks and measuring sticks is permitted: $g_{ik} \rightarrow \bar{g}_{ik} = \lambda(x) g_{ik}$ with a real scalar function λ . A physical meaning can be given to the class of all $\{\lambda g_{ik}\}$, only. The non-metricity tensor in this theory is $Q_{ij}{}^k = -g_{ij}Q^k$ with an arbitrary 1-form $Q = Q_k dx^k$. Upon re-gauging we have

$$Q_k \rightarrow \bar{Q}_k = Q_k + \frac{\partial(\log \lambda)}{\partial x^k} . \tag{14.10}$$

This equation invited Weyl to identify Q_k with the electromagnetic 4-potential. In his geometry, he called $dQ = (\frac{\partial Q_l}{\partial x^k} - \frac{\partial Q_k}{\partial x^l}) dx^k \wedge dx^l$ the *line-curvature* (Streckenkrümmung). It corresponded to the electromagnetic field.

In Weyl’s “gauge”theory of 1918 the norm of a tangent vector is *not* conserved under parallel transport along the closed curve \mathcal{C} . Instead

$$\left| \frac{dx^k}{du} \right| = \int_{\mathcal{C}} \exp \left[-Q_k(x(u)) \frac{dx^k}{du} \right] \tag{14.11}$$

⁶The name used in the Paris group for Equation (14.8), i.e., “equation de liaison” does not carry a physical interpretation.

holds. This means that the spectra of distant objects would depend on the path by which they reach the observer. Weyl's theory - to also be taken as a gauge theory for the Abelian gauge group $U(1)$ - thus did not form an acceptable geometrical model for UFT.⁷ For the history of gauge theory cf. O'Raifeartaigh and Straumann (2000).

For the description of curvature a *curvature tensor* is introduced; it may be defined by the non-commutativity of the 2nd covariant derivatives, or by parallel transport of a vector along the tangent vectors of a given closed curve. Its components are⁸:

$$K^l_{ijk} = K^l_{[ijk]} = \partial_j L_{ki}^l - \partial_k L_{ji}^l + L_{jn}^l L_{ki}^n - L_{kn}^l L_{ji}^n. \tag{14.12}$$

In general, two different *traces* of the curvature tensor exist: The so-called *Ricci-tensor* $K_{ij} =: K^l_{ijl}$ and the skew-symmetric *homothetical curvature*:

$$V_{ij} =: K^l_{lij} = \frac{\partial L_j}{\partial x^i} - \frac{\partial L_i}{\partial x^j} \tag{14.13}$$

with $L_j =: L_{jn}^n$.⁹ From the Ricci-tensor K_{ij} , in metric-affine or mixed geometry, a further trace may be formed, the so-called *curvature-* or *Ricci-scalar* $K =: g^{ik} K_{ik}(L)$. For an asymmetrical connection, torsion $S_{ki}^l =: L_{[ki]}^l$ has a geometrical meaning: it is proportional to the gap remaining when X is parallely transported along Y and Y along X .

The following *identity* among the 2-rank tensorial objects introduced exists:

$$V_{ij} + 2K_{[ij]} = 4S_{[i|j]} + S_{ij}^m{}_{|m} + 8S_{ij}^m S_m \tag{14.14}$$

where S_j is the so-called *vector torsion*.

In Einstein's gravitational theory L_{ij}^k agrees with the Christoffel- symbol $\{^k_{ij}\} =: \frac{1}{2}g^{(kl)}(g_{(li),j} + g_{(lj),i} - g_{(ij),l})$ fully determined by the Riemannian metric; V_{ij} , $K_{[ij]}$ and S_{ki}^l vanish identically. The Ricci scalar $K^l_l(g, \Gamma) \rightarrow R^l_l(g)$. While in general relativity, the Riemannian curvature tensor corresponds to the *gradients* of the gravitational field, in the fibre bundle formulation of gauge field theories the curvature tensor of the connection is related to the *gauge-fields* or field strenghts. We give a brief list of early attempted geometrizations in physics (Table 14.1).

⁷The gauge idea was successfully resurrected in 1928, however, within quantum mechanics. It later played a big role within non-abelian gauge theories.

⁸Note that alternative curvature tensors could and have been defined.

⁹We cannot discuss here the many different possibilities for building symmetric rank 2 tensors from the curvature tensor.

Table 14.1 Geometrizations of physics

Gravitation & inertia	Einstein & Grossmann(1913–1915)	<i>Riemann geometry, d=4</i>
Gravitation & electromagnetism	Weyl 1918	<i>Weyl geometry, d=4</i>
Gravitation & electromagnetism	Eddington 1921	<i>Affine geometry, d=4</i>
Gravitation & electromagnetism	Kaluza 1921 (1919)	<i>Riemann geometry, d=5</i>
Newtonian gravitation	Cartan, Weyl 1923, Friedrich 1927	<i>Metric-affine, d=1+3</i>
Gravitation & electromagnetism	Einstein 1925	<i>Mixed geometry, d=4</i>
Non-abelian gauge theories:	Electroweak interaction	<i>(Principal) fibre bundles</i>

14.3 Field Equations in Unified Field Theory, and How to Solve Them

While A. Einstein had been using mixed geometry in his work in UFT since the 1920s, Schrödinger entered the field only in 1943, within purely affine geometry. Two years later, Einstein also resumed his work—again within mixed geometry. His first three papers on UFT (Einstein 1925, 1945; Einstein and Straus 1946) all employ the metric and the connection as *independent* variables—with altogether 80 components in local coordinates while only 6 + 10 of them would be needed for a description of the gravitational and electromagnetic fields. In principle, some of the extra variables might be used to describe other fields and/or matter. In his approach, the symmetric part of the metric, i.e., h is taken to correspond to inertial and gravitational potentials while the skew-symmetric part k houses the electromagnetic field. The field equations are derived from such a Lagrangian that general relativity is contained in UFT as a *limiting* case. There are many possibilities for such a Lagrangian. Nevertheless, Einstein used a Lagrangian corresponding (more or less) to the curvature scalar in Riemannian geometry $\sqrt{-\det(g_{ik})} g^{lm} K_{lm}(L)$ without further justification. An additional term at hand would have been $\sqrt{-\det(g_{ik})} g^{lm} V_{lm}(L)$.

The field equations suggested, i.e., the *weak* Einstein-Straus equations, are:

$$L_i = 0 , \tag{14.15}$$

$$\hat{g}^{ik}_{||l} = 0 , \tag{14.16}$$

$$P_{(ik)} = 0 , \tag{14.17}$$

$$P_{[ik],l} + P_{[kl],i} + P_{[li],k} = 0 . \tag{14.18}$$

Here, $L_i := \frac{1}{2}(L_{im}^m - L_{mi}^m)$, $\hat{g}^{ik} := \sqrt{\det g_{lm}} g^{ik}$, and P_{ij} corresponds to one particular version of the Ricci-tensor, i.e., $P_{ij} = -\bar{K}^l{}_{ij}$. (Cf. section 2.1.3, eq. (23) of Goenner (2004)).

The first equation (14.15) was added by hand by Einstein. Equation (14.16) is equivalent to the compatibility condition (14.8). The system contains 82 equations for 80 variables. In the *strong* field equations, (14.17) and (14.18) are replaced by¹⁰

$$P_{ik} = 0. \quad (14.19)$$

The intention is to use (14.16) for expressing the affine connection in terms of the metric and of torsion, i.e., to completely remove the independent degrees of freedom in the connection.

Within affine geometry, Schrödinger arrived at a system of field equations in which (14.15) and (14.16) were retained but (14.17) and (14.18) replaced by

$$\begin{aligned} P_{(ik)} + \lambda g_{ik} = 0, & \quad (P_{[ik],l} + \lambda g_{[ik],l}) + (P_{[kl],i} \\ & \quad + \lambda g_{[kl],i}) + (P_{[li],k} + \lambda g_{[li],k}) = 0. \end{aligned} \quad (14.20)$$

Here $\lambda \neq 0$ is a constant comparable to the cosmological constant in general relativity. In his approach, the metric g_{ij} is a dependent quantity *defined* from a variational derivative of the Lagrangian with regard to P_{ik} . In later work, in particular by the Paris group of M.-A. Tonnelat, the name “Einstein-Schrödinger theory” was used often for the theory described by the weak field equations. In Figure 14.1, a table showing the periods during which some researchers worked in UFT is given.

It is here that we encounter ideology: H. Weyl’s idea of the connection to be taken seriously as an independent geometrical object and Eddington’s application of it in the form of a purely affine geometry occupied the mind of Einstein and of many of the other main players. For them, the first task was to solve (14.16) for the connection as a functional of the metric and its first derivative, i.e., to remove the connection from the field equations altogether and then to solve for the metric. This seemed necessary in order to arrive at *exact* solutions of the field equations (14.15)–(14.18). According to Einstein, the known particles should have been represented by singularity-free exact solutions corresponding to field concentrations. Otherwise, he could not have accepted the theory as being complete (Einstein 1952).

But why introduce an independent connection in the first place? The alternative, i.e., to directly construct a connection from the asymmetric metric, was not followed

¹⁰Thus 84 field equations hold for the 80 variables.

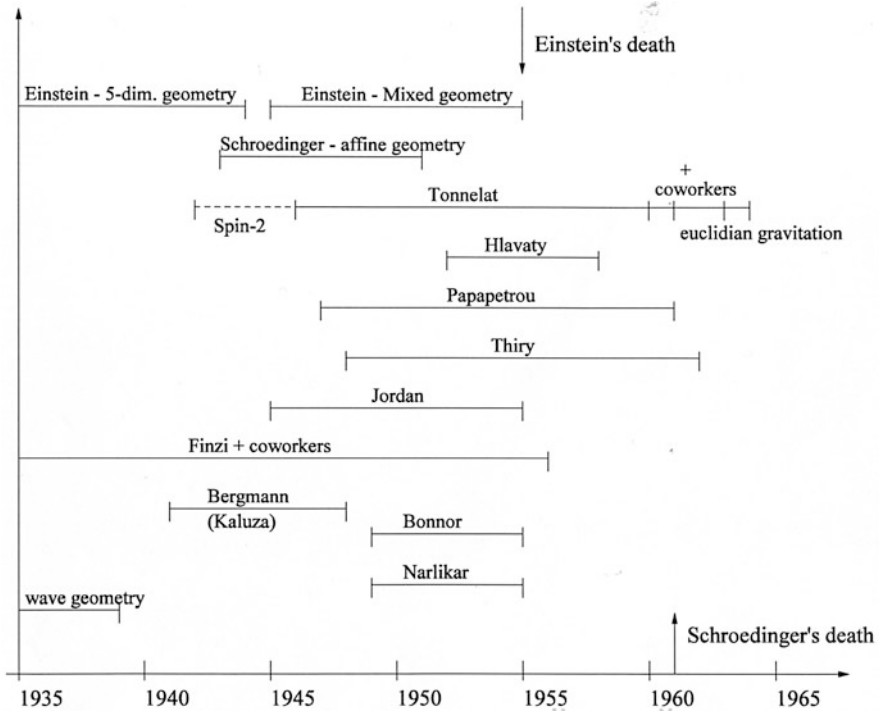


Fig. 14.1 Research periods of some workers in UFT

although already in 1928 K. Hattori had worked it out (Hattori 1928). Ten years earlier R. Bach alias Förster had tried to do the same and had been discouraged by Einstein (Schulmann et al. 1998, Doc. 439). Hattori’s connection is built from both the symmetric and antisymmetric parts of the metric¹¹:

$$\{..\}_{\text{Hattori}} := \frac{1}{2}h^{kr}(g_{ri, j} + g_{jr, i} - g_{ji, r}) . \tag{14.21}$$

By writing (14.21) in the form $\{..\}_{\text{Hattori}} := \{^k_{ij}\}_h + \frac{1}{2}h^{kr}(k_{ri, j} + k_{jr, i} - k_{ji, r})$, the decomposition (14.7) is reached. Hattori’s connection is not symmetric. Both, vector torsion and homothetic curvature are vanishing.¹² Thus, Schrödinger’s dictum, i.e.—“so much is certain, the structure of space-time negotiated by the g_{ik} -tensor, is exhausted. It yields gravitation and nothing more ” (Schrödinger 1946)—does

¹¹ $A_{, j} := \frac{\partial A}{\partial x^j}$.

¹² Note that another possibility would have been $\frac{1}{2}g^{kr}(g_{ri, j} + g_{jr, i} - g_{ji, r}) \equiv \{..\}_{\text{Hattori}} + 2k^{kl}h_{lm}\{^m_{ij}\}_h + \theta_{ij}{}^k$ with $\theta_{ij}{}^k = \frac{1}{2}k^{kr}(k_{ri, j} + k_{jr, i} - k_{ji, r})$.

not apply if the metric is asymmetric. It is also to be expected that the Ricci and curvature tensors house more degrees of freedom than the corresponding Riemannian objects. Thus, even the link between geometric objects and matter variables (electric 4-current) could be satisfied without making the detour of using a connection independent of the metric. In fact, for the electric 4-current j^k , among others the suggestions (Einstein 1950; Kursunoglu 1952):

$$k_{[ij],k} + k_{[ki],j} + k_{[jk],i} \quad (14.22)$$

and (Hlavatý 1958).

$$- \det(h_{lm}) \frac{\partial [h^{kr} h^{st} k_{rs}]}{\partial x^t}. \quad (14.23)$$

were made in mixed geometry. For their definition, a connection is not needed. On the other hand, also identifications of the electromagnetic field tensor with the skew-symmetric part of the Ricci tensor (Arnowitt 1957; Israel and Trollope 1961; Crumeyrolle 1962):

$$P_{[ik]}(L) \quad (14.24)$$

or with the homothetic curvature (Schrödinger 1946, 1948a,b):

$$V_{ik}(L) \quad (14.25)$$

were introduced. Instead of (14.22)

$$P_{[ij],k} + P_{[ki],j} + P_{[jk],i} \quad (14.26)$$

was also used for the electric 4-current (Einstein and Straus 1946; Israel and Trollope 1961).

14.3.1 How to Solve the “Compatibility” Equation

14.3.1.1 The Paris Groups: M.-A. Tonnelat, A. Lichnerowicz

The solution of the 64(= 40 + 24) compatibility equations (14.8) for the 64 components of the connection was very complicated. Einstein and Straus derived necessary and sufficient (algebraic) conditions for g_{ik} such that equations (14.8) determine the connection (in terms of the metric) “uniquely and without singularities”, but found no solution. Three years later, Straus tried again with partial success: he presented an implicit solution which was too unwieldy to be of any use (Straus 1949). It

Table 14.2 Paris research group of M.-A. Tonnelat. The *rightarrow* means that the person got his/her PhD from Mme. Tonnelat (M. de Broglie). The year of the dissertation is also shown

<i>L. de Broglie</i>	→ Marie-Antoinette Tonnelat (1941)
→ Jack Lévy (1957)	→ Pham Tan Hoang (1957)
→ Françoise Maurer née Tison (1958)	
→ S. Kichenassamy (1958)	→ Jean Hély (1959)
→ Marcel Bray (1960)	→ Liane Bouche née Valere (1961)
→ Marcel Lenoir (1962)	→ Philippe Droz-Vincent
→ Nguyen Phong Chau (1963)	→ Huyen-Dang-Vu
Stamatia Mavridès (1953)	Judith Winogradzki (1953)

was Marie-Antoinette Tonnelat who gave a general solution for the generic case in a series of four papers in 1949 and 1950 (Tonnelat 1949, 1950a,b,c). The trivial part known already from Riemann-Cartan theory is the observation that the affine connection is fully determined by the Christoffel symbol formed with the symmetric part of the metric and by the torsion tensor independent of the metric. The less trivial part is to solve the now available additional 24 equations for torsion as a functional of the metric and its 1st derivatives. Furthermore, the subtle question arose, whether (14.8) ought to be solved for the connection as a functional of the g_{ij} , or (14.16) as a functional of the g^{ij} . Stamatia Mavridès supplied the solution to the second alternative (Mavridès 1955). Certainly, these two solutions could always be recalculated from each other if just one of them was known. The real motivation behind the independent second calculation was to see whether h_{ij} should be taken as the metric for reasons of simplicity, or l^{ij} (l_{ik}) as Lichnerowicz had suggested. According to Mavridès, her calculation did not contribute to such a decision (Mavridès 1955, 1158–59). Table 14.2 shows some members of Tonnelat’s research group.

14.3.1.2 Attempts by Others

Also V. Hlavatý engaged himself in a systematic study of the problem and treated all possible cases, generic and degenerated (according to the rank of k_{ij}), in extenso (Hlavatý 1958). Four years after Mme. Tonnelat, S. N. Bose¹³ described a way of solving (14.8); his result depends on the calculation of eigenvectors and eigenvalues of certain tensors and is not apt to be used without further effort. He did not refer to any other publication (Bose 1954). In 1955, M.-A. Tonnelat (Tonnelat 1956) felt the need to secure priority for giving the solution of Equation (14.8), perhaps because

¹³It is the “Bose” from Bose-Einstein statistics.

S. N. Bose, in a paper in the journal of the French (!) Mathematical Society (Bose 1955), had improved on his earlier solution without mentioning Tonnelat's result.

In fact, of the four main figures in their work concerning UFT, Einstein and Schrödinger focussed on constructing acceptable and consistent field equations while Tonnelat and Hlavatý first concentrated on the solution of (14.8) and on drawing consequences from the field equations.

14.3.2 What Field Equations to Take?

Einstein already had had problems to arrive at the correct Maxwell's equations from the weak field equations of UFT: in linear approximation, he obtained equations much weaker than Maxwell's. They are given by:

$$\eta^{jk} \gamma_{[ij],k} = 0 ,$$

$$\eta^{mn} \partial_m \partial_n (\gamma_{[ik],l} + \gamma_{[kl],i} + \gamma_{[li],k}) = 0 .$$

Here, $g_{ij} = \eta_{ij} + \gamma_{ij} + O(\gamma^2)$. Moreover, after the first approximate calculations of the equations of motion for charged point particles tailored after the EIH-method (Einstein et al. 1938) had been made, it turned out that the electromagnetic field did not show up until and including the 4th order in the approximation scheme. Also, from the autoparallel equation, up to 2nd order approximation, only Newton's equations followed (Infeld 1950; Tonnelat 1955, 120–124). An escape from this untenable situation could have been that the approximation methods were not set up properly. Thus, Infeld's benevolent conclusion was:

“Yet it would not be right to regard this result as the death blow to Einstein's theory. The conclusions drawn here [...] follow along conventional lines as far as the electromagnetic field is concerned. It is possible that not Einstein's theory, but rather the conventional approach is at fault.” (Infeld 1950, 284)

Nevertheless, after the first *exact* spherically symmetric and static solutions of the “weak” and “strong” field equations had been found, the situation remained the same: it became more and more dubious whether the weak field equations could describe a field as simple as the electromagnetic field of a slowly moving point charge.

For this reason, very soon other field equations were suggested. M. A. Tonnelat started from a modification of Einstein's weak field equations:

$$\hat{g}_{ik}^{+-} |l = 0; \partial_l \hat{F}^{il} = 0; K_{ik}(\tilde{L}) = 0, \quad (14.27)$$

$$K_{[ij],k}(\tilde{L}) + K_{[ki],j}(\tilde{L}) + K_{[jk],i}(\tilde{L}) = 0, \quad (14.28)$$

where the covariant derivative and the Ricci tensor are formed with regard to a connection with vanishing vector torsion $\tilde{L}_{ij}^k =: L_{ij}^k + \frac{2}{3} \delta_i^k L_j$ with $L_i =: L_{[il]}$.

Thus $\tilde{L}_i =: \tilde{L}_{[il]}^l = 0$. Moreover, $\hat{F}^{ik} =: \sqrt{-g} f^{ik}$ with f^{ik} being the matrix reciprocal to k_{ik} : $k_{il} f^{jl} = \delta_i^j$. Equations (14.27), (14.28) follow from a variational principle with Lagrangian $\mathcal{L} = \hat{g}^{ik} K_{ik}(\tilde{L})$ (Tonnelat 1955, 31).

A student of Mme. Tonnelat, Francoise Maurer, née Tison, in 1959 used the following field equations (Maurer-Tison 1959, 203):

$$\partial_l g_{ik} - \tilde{L}_{il}^m g_{mk} - \tilde{L}_{lk}^m g_{im} = 0, \tag{14.29}$$

$$\partial_l (g^{[lk]} \sqrt{-g}) = 0, \tag{14.30}$$

$$P_{ik}(\tilde{L}) - \frac{2}{3}(\partial_i L_k - \partial_k L_i) = 0, \tag{14.31}$$

where L_{ik}^l is a linear connection with torsion vector L_k and \tilde{L}_{ik}^l a linear connection with vanishing torsion vector $\tilde{L}_k = 0$. $P_{ik}(\tilde{L})$ is the Ricci tensor of the connection \tilde{L}_{il}^m .¹⁴

W. B. Bonnor kept (14.15) and (14.16) but suggested the other equations to be supplied with an additional term (Bonnor 1954, 1957):

$$P_{(ik)} + p^2 U_{ik} = 0, \tag{14.32}$$

$$P_{[ik],l} + P_{[kl],i} + P_{[li],k} + p^2(U_{[ik],l} + U_{[kl],i} + U_{[li],k}) = 0, \tag{14.33}$$

where p is an arbitrary constant and U_{ik} is defined by:

$$U_{ik} = g_{[ki]} - g^{[st]} g_{is} g_{tk} + \frac{1}{2} g^{[st]} g_{st} g_{ik}. \tag{14.34}$$

Again following the EIH-approximation procedure, he succeeded to derive, in 4th order, the equation of motion for two charged, massive point particles - one fixed in $\vec{r} = 0$, the other slowly moving:

$$m_1 \frac{d^2 \vec{r}}{dt^2} = -\frac{m_1 m_2 \vec{r}}{r^3} + p^2 q^2 \frac{e_1 e_2 \vec{r}}{r^3}. \tag{14.35}$$

¹⁴Maurer-Tison introduced

$$g_{\alpha\beta} = h_{\alpha\beta} + k_{\alpha\beta}.$$

$$g^{\alpha\beta} = l^{\alpha\beta} + m^{\alpha\beta}.$$

Here, m_1, m_2 are the masses, e_1, e_2 the electric charges, and \vec{r} the relative distance.¹⁵ Note that geodesics and autoparallels no longer coincide as in Riemannian geometry. This leads to a further ambiguity.

14.4 Persons, Publications, Interactions

14.4.1 Persons, Places and Publications

We first note the places at which research in UFT mainly went on. As mentioned four of them were particularly important: Princeton (Einstein), Dublin (Schrödinger), Paris (Mme. Tonnelat, A. Lichnerowicz) and less so Bloomington, Indiana (Hlavaty - working mostly by himself). All in all, about 150–170 scientists did take part in research on UFT between 1930 and 1965 with a yearly average of 18 papers published.¹⁶ The maximum of a 5-year-average in published papers per year at 40 papers occurred in the years just before and just after Einstein's death, i.e., in 1953–1957; the minimum with 4 papers in the period 1939–1943, during the war years (cf. Figure 14.2). We also look at the more detailed map showing 3

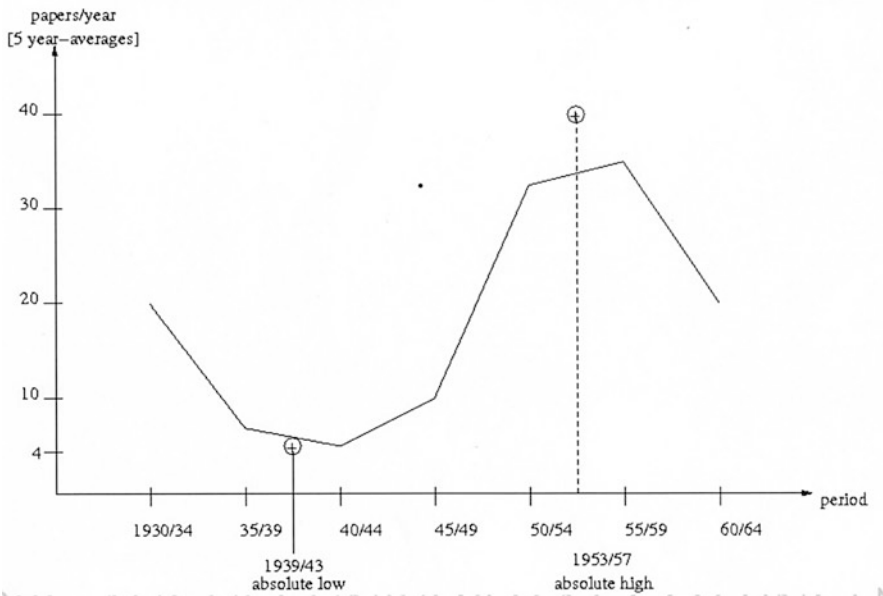


Fig. 14.2 Five-year-average of total publications per year on UFT

¹⁵ q is another constant.

¹⁶At present, my databank contains 632 papers from 1930–1965.

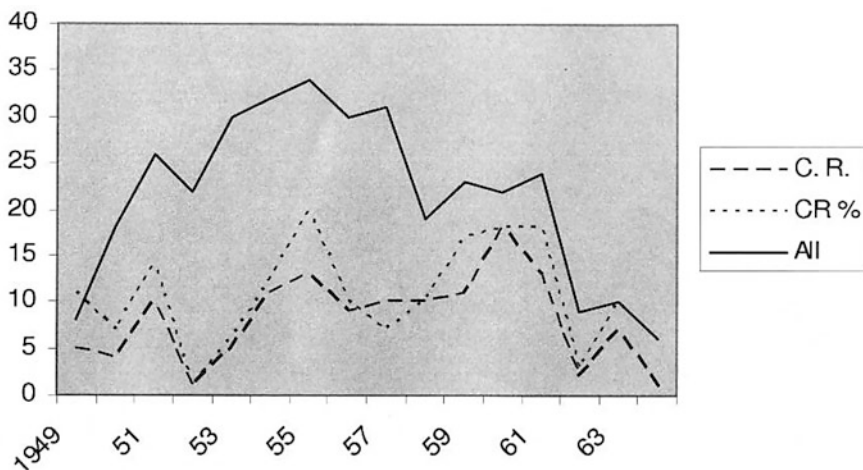


Fig. 14.3 Yearly publications on UFT. Solid line: all publications (my database); dashed line: absolute number of papers on UFT in *Comptes rendus*; dotted line: percentage of papers on UFT in *Comptes rendus* compared to papers on General Relativity in the same Journal

plots: (1) absolute numbers of total yearly publications in UFT ; (2) absolute number of those published in “*Comptes Rendus de l’Académie des Sciences, Paris*”, and (3) their percentage relative to the papers on general relativity likewise published in “*Comptes Rendus*”. The number of papers on UFT at most reached 20% of the papers on general relativity (cf. Figure 14.3). At the beginning of the 1960s most scientists working in UFT dropped the subject in favour of general relativity or alternative theories of *gravitation*. Research in “unified field theory” quickly received the label of being an outmoded subject.

14.4.2 Interactions

In order to obtain an impression of how people in the field interacted, several questions may be asked: Are there co-authorships, inside and outside the various groups? Mutual visits or even changes in permanent position? Who cites whom? Who acts as a referee to whom? Correspondence? The investigation of the last item being the most time- (and money-) consuming is outside my reach. Think of Einstein’s large correspondence; even Hlavatý corresponded with ca. 40 scientists working on UFT. As concerns the other questions, I shall make only a few remarks.

As to co-authorships, it was not to be expected that some of the group leaders, senior professors of the old style, would join to write a paper. This did not even occur between “colleagues” Marie-Antoinette Tonnelat and André Lichnerowicz in Paris. As co-authors, we most often have the combination of group leader and

coworker/PhD-student: Einstein & Straus, Einstein & Kaufmann etc., Schrödinger & Mautner, Schrödinger & Papapetrou etc, Marie-Antoinette Tonnelat & Liane Bouche, Tonnelat & Sylvie Lederer; Lichnerowicz & Thiry, Lichnerowicz & Fourès-Bruhat, Hlavatý & Saenz. Remarkably, Tonnelat did not publish jointly with her coworkers Judith Winogradzki (née Winterberg) (1916–2006) and Stamatia Mavridès who both had not obtained their PhD with her. Winogradzki worked on UFT from 1954–1956 and then on spinors. Mavridès on UFT from 1954–1957, and on Tonnelat’s euclidean theory of gravity from 1962–1964. Thereafter, she went into cosmology and astrophysics. In 1971 Tonnelat and Mavridès visited Brasil (Rio de Janeiro) as guest professors. Table 14.3 shows Einstein’s collaborators in Princeton.

Did they move around much? During their work on UFT, the group leaders Einstein, Schrödinger, Lichnerowicz, and Tonnelat (with the exception of her short sojourn in Dublin) stayed at the same place. Hlavatý spent some time in Princeton following an invitation by Einstein, and in Paris as a guest professor (1948) after he had left Czechoslovakia and before he came to Indiana University. A. Papapetrou (1907–1997) who had been a professor in Greece during 1940–1946, was the only one contributing to UFT who really moved around. He worked in Dublin until 1948, at the University of Manchester until 1952 when he went to Berlin (G.D.R.). There, he lead a group in general relativity at the Academy of Sciences until 1962. From then on he stayed in Paris until retirement as “Directeur de recherche” at the CNRS.

As to mutual referencing, I have only looked at the Italian group around Finzi, Pastori and Udeschini which was left in the background up to here. (For some of the researchers involved see Table 14.4.) In their publications, references to work inside the group are given 55 times and to outside papers 82 times.

Table 14.3 Princeton group.
Straus and Kaufmann did
take part in UFT proper

Albert Einstein
Nathan Rosen [1934–1935]
Leopold Infeld [1936–1937]
Banesh Hoffmann [1935–1937]
Valentin Bargmann [1937–1946]
Peter Bergmann [1936–1941]
Ernst Straus [1950–1953]
Bruria Kaufmann [1950–1955]

Table 14.4 Scientists working on UFT in Italy

Bruno Finzi (Milano)	Maria Pastori (Milano)
→ Paulo Udeschini (Pavia)	→ Emilio Clauser (Milano)
→ Elisa Brinis-Udeschini (Milano)	→ Laura Gotusso (Milano)
→ Bartolomeo Todeschini (Milano)	→ Franco De Simoni (Pisa)
→ Franca Graiff (Milano)	→ Laura Martuscelli
→ Angelo Zanella (Milano)	→ [Luigia Mistrangioli]

From the research done outside of Italy, Princeton (Einstein, AE & Straus, AE & Kaufman) is quoted most often (24 times); next in frequency are Hlavatý's papers (referred to 18 times), then Schrödinger's and Papapetrou's research in Dublin (quoted 11 times). The remaining citations are distributed geographically almost equally: USA/Canada (7 citations) India (6), Japan and Paris (Lichnerowicz, Tonnelat) 5 each, and England (Bonnor, Stephenson) (4 citations). Thus, the contact to those closest at hand geographically, i.e., the Paris group, was almost negligible. This was reciprocated by Tonnelat: in her book (Tonnelat 1955), she refers to the Italian groups nine times, whereas Einstein and his collaborators are listed 28 times. Next follows Schrödinger with 21 bibliographic entries and Hlavatý with 6. She lists 21 references to papers by herself or by members of her group.

The reports in "Mathematical Reviews" are signed by the reviewers; mostly they just describe the papers' contents. Sometimes, however they did not hold back criticism. Thus the reviews mainly reveal public visibility of UFT and, much less so, a judgment on the quality of the research. My study of 133 reports written by 29 reviewers in "Mathematical Reviews" shows that only 40% of the reviewers were involved in research in UFT themselves. Einstein, Schrödinger and Tonnelat have not written reports. The applied mathematician A. H. Taub who made valuable contributions to general relativity alone provided about 23% of the reviews. He concentrated on papers by Tonnelat and Schrödinger. The next largest set of reviews was written by Hlavatý who dealt with papers from the Paris groups (Tonnelat, Lichnerowicz) and his Princeton colleague Eisenhart. There is some reviewing even between the Paris "subgroups": of Mme Tonnelat by Mme. Choquet-Fourès(-Bruhat), of Mme Maurer-Tison and Jacques Lévy by Mme. Renaudie.

14.5 Conclusion

The chords struck by Hermann Minkowski in 1908 and Albert Einstein in 1915 still ring today; and with them the ideology of *unification in the form of geometrization* of fields and matter. It seems that this idea was luring most of the researchers into UFT. To the more mathematically interested minds, the new challenges posed by non-linear field equations and geometries more general than the Riemannian might have also provided a motivation. Above all, the towering personality of Albert Einstein along with Nobel laureate Erwin Schrödinger, with their common dislike for quantum mechanics, stabilized the community. Once both had died, by the beginning of the 1960s the UFT-community quickly dispersed and its members chose new fields of interest. Since the mid 50s, Yang-Mills theory, using non-abelian gauge fields, came to light and proved to be very successful. While UFT never was able to become a "mainstream"-subject, Yang-Mills theory did. Nevertheless, a few aficionados carried on with classical UFT (mixed geometry) in the 70s and 80s - with some following the original direction of this research even to this day.

References

- Arnowitz, R. L. (1957). Phenomenological approach to a unified field theory. *Physical Review*, 105, 735–742.
- Bonnor, W. B. (1954). The equation of motion in the non-symmetric unified field theory. *Proceedings of the Royal Society of London, A* 226, 366–377.
- Bonnor, W. B. (1957). Les équations du mouvement en théorie unitaire d'Einstein-Schrödinger. *Annales de l'Institut H. Poincaré.*, 15, fasc. 3, 133–145.
- Bose, S. (1954). The affine connection in Einstein's new unitary field theory. *Annals of Mathematics*, 59, 171–176.
- Bose, S. (1955). Solution d'une équation tensorielle intervenant dans la théorie du champ unitaire. *Bulletin de la Société Mathématique de France*, 83, 81–88.
- Crumeyrolle, A. (1962). Sur quelques interprétations physiques et théoriques des équations du champ unitaire d'Einstein-Schrödinger. *Rivista di matematica della Università di Parma* (2), 3, 331–391.
- Einstein, A. (1925). Einheitliche Feldtheorie von Gravitation und Elektrizität. *Sitzungsberichte der Preussischen Akademie der Wissenschaften*, 22, 414–419.
- Einstein, A. (1945). A Generalization of the relativistic theory of gravitation. *Annals of Mathematics*, 46, 578–584.
- Einstein, A. (1950). *The meaning of relativity* (4th edn.). London: Methuen. Appendix II: "Generalization of Gravitation Theory."
- Einstein, A. (1952). A comment on a criticism of unified field theory. *Physical Review*, 89, 321.
- Einstein, A., Infeld, L., Hoffmann, B. (1938). The gravitational equations and the problem of motion. *Annals of Mathematics*, 39, 65–100.
- Einstein, A., & Straus, E. G. (1946). A Generalization of the relativistic theory of gravitation. II. *Annals of Mathematics*, 47, 731–741.
- Goenner, H. (2004). On the history of unified field theory. *Living Reviews in Relativity*, 7, 2. <https://doi.org/10.12942/lrr-2004-2>.
- Goenner, H. (2014). On the history of unified field theory, Part II. *Living Reviews in Relativity*, 17, 5. <https://doi.org/10.12942/lrr-2014-5>.
- Hattori, K. (1928). Über eine formale Erweiterung der Relativitätstheorie und ihren Zusammenhang mit der Theorie der Elektrizität. *Physikalische Zeitschrift*, 29, 538–549.
- Hlavatý, V. (1958). *Geometry of Einstein's unified field theory*. Groningen: P. Noordhoff.
- Infeld, L. (1950). The new Einstein theory and the equations of motion. *Acta Physica Polonica*, 10, fasc.3–4, 284–293.
- Israel, W., & Trollope, R. (1961). New possibilities for a unified field theory. *Journal of Mathematical Physics*, 2, 777–786.
- Kursunoglu, B. (1952). Gravitation and electrodynamics. *Physical Review*, 88, 1369–1379.
- Licherowicz, A. (1953). Compatibilité des Équations de la Théorie Unitaire d'Einstein-Schrödinger. *Comptes Rendus de l'Académie des Sciences, Paris*, 237, 1383–1386.
- Licherowicz, A. (1954). Compatibilité des Équations de la Théorie Unitaire du Champ d'Einstein. *Journal of Rational Mechanics and Analysis*, 3, 487–521.
- Licherowicz, A. (1955a). *Les théories relativistes de la gravitation et de l'électromagnétisme*. Paris: Masson.
- Licherowicz, A. (1955b). Étude des Équations de la Théorie Unitaire d'Einstein. *Rendiconti del Seminario Matematico e Fisico di Milano*, 25, 121–133.
- Maurer-Tison, F. (1959). Aspects mathématiques de la théorie unitaire du champ d'Einstein. *Annales Scientifiques de l'école Normale Supérieure*, 3e série, t. 76, 185–269.
- Mavridès, S. (1955). La solution générale des équations d'Einstein $g^{+\mu\nu;\rho} = 0$. *Nuovo Cimento*, 2, Series 10, 1141–1164.
- O'Raifeiraigh, L., & Straumann, N. (2000). Gauge theory: Historical origins and some modern developments. *Reviews of Modern Physics*, 72, 1–23.

- Schrödinger, E. (1946). Affine feldtheorie und meson. *Verhandlungen der Schweizerischen Naturforschenden Gesellschaft*, 126, 53–61.
- Schrödinger, E. (1948a). The final affine field laws. II. *Proceedings of the Royal Irish Academy*, A51, 205–216.
- Schrödinger, E. (1948b). The final affine field laws III. *Proceedings of the Royal Irish Academy*, 52A, 1–9.
- Schulmann, R., Kox A. J., Janssen, M., & Illy, J. (Eds.) (1988). *The collected papers of Albert Einstein* (Vol. 8). Princeton: Princeton University Press.
- Straus, E. G. (1949). Some results in Einstein's unified field theory. *Review of Modern Physics*, 21, 414–420.
- Tonnellat, M.-A. (1949). Théorie unitaire du champ physique. 1. Les tenseurs fondamentaux et la connexion affine. 2. Cas d'une métrique symétrique. 3. Détermination des tenseurs fondamentaux. *Comptes rendus de l'academie des Sciences*, 228, 368–370, 660–662, 1846–1848.
- Tonnellat, M.-A. (1950a). Résolution des equations fondamentales d'une théorie unitaire affine. *Comptes rendus de l'academie des Sciences*, 230, 182–184.
- Tonnellat, M.-A. (1950b). Théorie unitaire affine. I. Choix des tenseurs de base et obtension de l'équation fondamentale. *Comptes rendus de l'academie des Sciences*, 231, 470–472.
- Tonnellat, M.-A. (1950c). Théorie unitaire affine. II. Résolution rigoureuse de l'équation fondamentale. *Comptes rendus de l'academie des Sciences*, 231, 487–489.
- Tonnellat, M.-A. (1955). *La théorie du champ unifié d'Einstein et quelques-uns de ses développements*. Paris: Gauthier Villars.
- Tonnellat, M.-A. (1956). La solution générale des équations d'Einstein $g_{\mu\nu} = 0$. In A. Mercier & M. Korvaire (Eds.), *Jubilee of Relativity Theory. Bern, 11–16 July 1955* (pp. 192–197). Basel: Birkhäuser.
- Weyl, H. (1919). Gravitation und Elektrizität. *Sitzungsberichte der Preussischen Akademie der Wissenschaften*, Nr. 26, 465–478 with a “Nachtrag” by Einstein, p. 478, and “Erwiderung des Verfassers”, pp. 478–480.

Part V
**Quantum Gravity, Conformal Boundaries,
and String Theory**

Chapter 15

The Formulation of Quantum Field Theory in Curved Spacetime



Robert M. Wald

Quantum field theory in curved spacetime is a theory wherein matter is treated fully in accord with the principles of quantum field theory, but gravity is treated classically in accord with general relativity. It is not expected to be an exact theory of nature, but it should provide a good approximate description in circumstances where the quantum effects of gravity itself do not play a dominant role. Despite its classical treatment of gravity, quantum field theory in curved spacetime has provided us with some of the deepest insights we presently have into the nature of quantum gravity.

Quantum field theory as usually formulated contains many elements that are very special to Minkowski spacetime. But we know from general relativity that spacetime is not flat, and, indeed there are very interesting quantum field theory phenomena that occur in contexts—such as in the early universe and near black holes—where spacetime cannot be approximated as nearly flat.

It is a relatively simple matter to generalize classical field theory from flat to curved spacetime. That is because there is a clean separation between the field equations and the solutions. The field equations can be straightforwardly generalized to curved spacetime in an entirely local and covariant manner. Solutions to the field equations need not generalize from flat to curved spacetime, but this doesn't matter for the formulation of the theory.

In quantum field theory, “states” are the analogs of “solutions” in classical field theory. However, properties of states—in particular, the existence of a Poincaré invariant vacuum state—are deeply embedded in the usual formulations of quantum field theory in Minkowski spacetime. For this reason and a number of other reasons,

R. M. Wald (✉)

Enrico Fermi Institute and Department of Physics, University of Chicago,
5640 S. Ellis Avenue, Chicago, IL 60637, USA
e-mail: rmwa@uchicago.edu

it is highly nontrivial to generalize the formulation of quantum field theory from flat to curved spacetime.

As a very simple, concrete example of a quantum field theory that illustrates some key features of quantum field theory as well as some of the issues that arise in generalizing the formulation of quantum field theory to curved spacetime, consider a free, real Klein-Gordon field in flat spacetime,

$$\partial^a \partial_a \phi - m^2 \phi = 0 . \quad (15.1)$$

The usual route towards formulating a quantum theory of ϕ is to decompose it into a series of modes, and then treat each mode by the rules of ordinary quantum mechanics. To avoid technical awkwardness, it is convenient to imagine that the field is in cubic box of side L with periodic boundary conditions. We can then decompose ϕ as a Fourier series in terms of the modes

$$\phi_{\mathbf{k}} \equiv L^{-3/2} \int e^{-i\mathbf{k}\cdot\mathbf{x}} \phi(t, \mathbf{x}) d^3x \quad (15.2)$$

where $\mathbf{k} = \frac{2\pi}{L}(n_1, n_2, n_3)$. The Hamiltonian of the system is then given by

$$H = \sum_{\mathbf{k}} \frac{1}{2} \left(|\dot{\phi}_{\mathbf{k}}|^2 + \omega_{\mathbf{k}}^2 |\phi_{\mathbf{k}}|^2 \right) \quad (15.3)$$

where

$$\omega_{\mathbf{k}}^2 = |\mathbf{k}|^2 + m^2 . \quad (15.4)$$

Thus, a free Klein-Gordon field, ϕ , in flat spacetime is explicitly seen to be simply an infinite collection of decoupled harmonic oscillators. If we take into account the fact that ϕ is real but the modes $\phi_{\mathbf{k}}$ are complex, we find that each $\phi_{\mathbf{k}}$ is given by the operator expression

$$\phi_{\mathbf{k}} = \frac{1}{2\omega_{\mathbf{k}}} (a_{\mathbf{k}} + a_{-\mathbf{k}}^\dagger) . \quad (15.5)$$

where $a_{\mathbf{k}}$ and $a_{\mathbf{k}}^\dagger$ satisfy the usual commutation relations

$$[a_{\mathbf{k}}, a_{\mathbf{k}'}] = 0 , \quad [a_{\mathbf{k}}, a_{\mathbf{k}'}^\dagger] = \delta_{\mathbf{k}\mathbf{k}'} I \quad (15.6)$$

The Heisenberg field operator $\phi(t, \mathbf{x})$ is then formally given by

$$\phi(t, \mathbf{x}) = L^{-3/2} \sum_{\mathbf{k}} \frac{1}{2\omega_{\mathbf{k}}} \left(e^{i\mathbf{k}\cdot\mathbf{x} - i\omega_{\mathbf{k}}t} a_{\mathbf{k}} + e^{-i\mathbf{k}\cdot\mathbf{x} + i\omega_{\mathbf{k}}t} a_{\mathbf{k}}^\dagger \right) . \quad (15.7)$$

However, this formula does *not* make sense as a definition of ϕ as an operator at each point (t, \mathbf{x}) . In essence, the contributions from the modes at large $|\mathbf{k}|$ do not diminish rapidly enough with $|\mathbf{k}|$ for the sum to converge. However, these contributions are rapidly varying in spacetime, so if we “average” the right side of Equation (15.7) in an appropriate manner over a spacetime region, the sum will converge. More precisely, Equation (15.7) defines ϕ as an “operator valued distribution”, i.e., for any (smooth, compactly supported) “test function”, f , the quantity

$$\phi(f) = \int f(t, \mathbf{x})\phi(t, \mathbf{x})d^4x \quad (15.8)$$

is well defined by Equation (15.7) if the integration is done prior to the summation.

States of the free Klein-Gordon field are given the following interpretation: The state, denoted $|0\rangle$, in which all of the oscillators comprising the Klein-Gordon field are in their ground state is interpreted as representing the “vacuum”. States of the form $(a^\dagger)^n|0\rangle$ are interpreted as ones where a total of n “particles” are present.

In an interacting theory, the state of the field may be such that the field behaves like a free field at early and late times. In that case, one has a particle interpretation at early and late times. The relationship between the early and late time particle descriptions of a state—given by the S-matrix—contains a great deal of the dynamical information about the interacting theory.

The particle interpretation/description of quantum field theory in flat spacetime has been remarkably successful—to the extent that one might easily get the impression that, at a fundamental level, quantum field theory is really a theory of “particles”. However, note that even for a free field, the definition and interpretation of the “vacuum” and “particles” depends heavily on the ability to decompose the field into its positive and negative frequency parts (as can be seen explicitly from Equation (15.7) above). The ability to define this decomposition makes crucial use of the presence of a time translation symmetry in the background Minkowski spacetime. In a generic curved spacetime without symmetries, there is no natural notion of “positive frequency solutions” and, consequently, no natural notion of a “vacuum state” or of “particles”.

If one looks more deeply at the usual general formulations of quantum field theory, it can be seen that many other properties that are special to Minkowski spacetime are used in an essential way. This is well illustrated by examining the Wightman axioms, since these axioms abstract the key features of quantum field theory in Minkowski spacetime in a mathematically clear way. We will focus attention on the Wightman axioms below, but a similar discussion would apply to other approaches, including the much less rigorous textbook treatments of quantum field theory.

The Wightman axioms are the following (Streater and Wightman 1964):

- The states of the theory are unit rays in a Hilbert space, \mathcal{H} , that carries a unitary representation of the Poincaré group.
- The 4-momentum (defined by the action of the Poincaré group on the Hilbert space) is positive, i.e., its spectrum is contained within the closed future light cone (“spectrum condition”).

- There exists a unique, Poincaré invariant state (“the vacuum”).
- The quantum fields are operator-valued distributions defined on a dense domain $\mathcal{D} \subset \mathcal{H}$ that is both Poincaré invariant and invariant under the action of the fields and their adjoints.
- The fields transform in a covariant manner under the action of Poincaré transformations.
- At spacelike separations, quantum fields either commute or anticommute.

It is obvious that there are serious difficulties with extending the Wightman axioms to curved spacetime, specifically:

- A generic curved spacetime will not possess any symmetries at all, so one certainly cannot require “Poincaré invariance/covariance” or invariance under any other type of spacetime symmetry.
- Even for a free quantum field, there exist unitarily inequivalent Hilbert space constructions of the theory. For spacetimes with a noncompact Cauchy surface—and in the absence of symmetries of the spacetime—none appears “preferred”.
- In a generic curved spacetime, there is no “preferred” choice of a “vacuum state”.
- There is no analog of the spectrum condition in curved spacetime that can be formulated in terms of the “total energy-momentum” of the quantum field.

Thus, of all of the Wightman axioms, only the last one (commutativity or anticommutativity at spacelike separations) generalizes straightforwardly to curved spacetime.

I will now explain in more detail some of the difficulties associated with generalizing the spectrum condition and the existence of a preferred vacuum state to curved spacetime:

Total Energy in Curved Spacetime The stress energy tensor, T_{ab} , of a classical field in curved spacetime is well defined and it satisfies local energy-momentum conservation in the sense that $\nabla^a T_{ab} = 0$. If t^a is a vector field on spacetime representing time translations and Σ is a Cauchy surface, one can define the total energy, E , of the field at “time” Σ by

$$E = \int_{\Sigma} T_{ab} t^a n^b d\Sigma . \quad (15.9)$$

Classically, for physically reasonable fields, the stress-energy tensor satisfies the dominant energy condition, so $T_{ab} t^a n^b \geq 0$. Thus, classically, we have $E \geq 0$. However, unless t^a is a Killing field (i.e., unless the spacetime is stationary), E will not be conserved, i.e., independent of choice of Cauchy surface, Σ .

In quantum field theory, it is expected that the stress-energy operator will be well defined as an operator-valued distribution, and it is expected to be conserved, $\nabla^a T_{ab} = 0$; see Hollands and Wald (2005). However, the definition of T_{ab} requires spacetime smearing. In Minkowski spacetime, since E is conserved one can, in effect, do “time smearing” without changing the value of E , and there is a unique, well defined notion of total energy. However, in the absence of time translation

symmetry, one cannot expect E to be well defined at a “sharp” moment of time. More importantly, it is well known that T_{ab} cannot satisfy the dominant energy condition in quantum field theory (even if it holds for the corresponding classical theory); locally, energy densities can be arbitrarily negative. It is nevertheless true in Minkowski spacetime that the total energy is positive for physically reasonable states. However, in a curved spacetime without symmetries, there is no reason to expect any “time smeared” version of E to be positive. Furthermore, there are simple examples with time translation symmetry (such as a two-dimensional massless Klein-Gordon field in an $S^1 \times \mathbf{R}$ universe) where E can be computed explicitly and is found to be negative (Birrell and Davies 1982). Thus, it appears hopeless to generalize the spectrum condition to curved spacetime in terms of the positivity of a quantity representing “total energy”.

Nonexistence of a “Preferred Vacuum State” and a Notion of “Particles” As already noted above, for a free field in Minkowski spacetime, the notion of “particles” and “vacuum” is intimately tied to the notion of “positive frequency solutions”, which, in turn relies on the existence of a time translation symmetry. These notions of a (unique) “vacuum state” and “particles” can be straightforwardly generalized to (globally) stationary curved spacetimes. However, there is no natural notion of “positive frequency solutions” in a general, nonstationary curved spacetime.

Nevertheless for a free field on a general curved spacetime, a notion of “vacuum state” can be defined as follows. A state is said to be *quasi-free* if all of its n -point correlation functions $\langle \phi(x_1) \dots \phi(x_n) \rangle$ can be expressed in terms of its 2-point correlation function by the same formula as holds for the ordinary vacuum state in Minkowski spacetime. A state is said to be *Hadamard* if the singularity structure of its 2-point correlation function $\langle \phi(x_1)\phi(x_2) \rangle$ in the coincidence limit $x_1 \rightarrow x_2$ is the natural generalization to curved spacetime of the singularity structure of $\langle 0|\phi(x_1)\phi(x_2)|0 \rangle$ in Minkowski spacetime (see Equation (15.13) below). Thus, in a general curved spacetime, the notion of a quasi-free Hadamard state provides a notion of a “vacuum state”, associated to which is a corresponding notion of “particles”. The problem is that this notion of vacuum state is highly non-unique. Indeed, for spacetimes with a non-compact Cauchy surface, different choices of quasi-free Hadamard states give rise, in general, to unitarily inequivalent Hilbert space constructions of the theory, so in this case it is not even clear what the correct Hilbert space of states should be. In the absence of symmetries or other special properties of a spacetime, there does not appear to be any preferred choice of quasi-free Hadamard state.

In my view, the quest for a “preferred vacuum state” in quantum field theory in curved spacetime is much like the quest for a “preferred coordinate system” in classical general relativity. After our more than 90 years of experience with classical general relativity, there is a consensus that it is fruitless to seek a preferred coordinate system for general spacetimes, and that the theory is best formulated geometrically, wherein one does not have to specify a choice of coordinate system in order to formulate the theory. Similarly, after our more than 40 years of experience with quantum field theory in curved spacetime, it seems similarly clear to me that it is fruitless to seek a preferred vacuum state for general spacetimes, and that the

theory should be formulated in a manner that does not require one to specify a choice of state (or representation) to define the theory.

Nevertheless, many of the above difficulties can be resolved in an entirely satisfactory manner:

- The difficulties that arise from the existence of unitarily inequivalent Hilbert space constructions of quantum field theory in curved spacetime can be overcome by formulating the theory via the algebraic framework. The algebraic approach also fits in very well with the viewpoint naturally arising in quantum field theory in curved spacetime that the fundamental observables in quantum field theory are the local quantum fields themselves.
- The difficulties that arise from the lack of an appropriate notion of the total energy of the quantum field can be overcome by replacing the spectrum condition by a “microlocal spectrum condition” that restricts the singularity structure of the expectation values of the correlation functions of the local quantum fields in the coincidence limit.
- Many aspects of the requirement of Poincaré invariance of the quantum fields can be replaced by the requirement that the quantum fields be locally and covariantly constructed out of the metric.

I will now explain these resolutions in more detail:

The Algebraic Approach In the algebraic approach, instead of starting with a Hilbert space of states and then defining the field observables as operators on this Hilbert space, one starts with a $*$ -algebra, \mathcal{A} , of field observables. A *state*, ω , is simply a linear map $\omega : \mathcal{A} \rightarrow \mathbf{C}$ that satisfies the positivity condition $\omega(A^*A) \geq 0$ for all $A \in \mathcal{A}$. The quantity $\omega(A)$ is interpreted as the expectation value of the observable A in the state ω .

If \mathcal{H} is a Hilbert space which carries a representation, π , of \mathcal{A} , and if $\Psi \in \mathcal{H}$ then the map $\omega : \mathcal{A} \rightarrow \mathbf{C}$ given by

$$\omega(A) = \langle \Psi | \pi(A) | \Psi \rangle \quad (15.10)$$

defines a state on \mathcal{A} in the above sense. Conversely, given a state, ω , on \mathcal{A} , we can use it to obtain a Hilbert space representation of \mathcal{A} by the following procedure, known as the Gelfand-Naimark-Segal (GNS) construction. First, we use ω to define a (pre-)inner-product on \mathcal{A} by

$$(A_1, A_2) = \omega(A_1^* A_2). \quad (15.11)$$

By factoring by zero-norm vectors and completing this space, we get a Hilbert space \mathcal{H} , which carries a natural representation, π , of \mathcal{A} . The vector $\Psi \in \mathcal{H}$ corresponding to $I \in \mathcal{A}$ then satisfies $\omega(A) = \langle \Psi | \pi(A) | \Psi \rangle$ for all $A \in \mathcal{A}$.

Thus, the algebraic approach is not very different from the usual Hilbert space approach in that every state in the algebraic sense corresponds to a state in the Hilbert space sense and vice-versa. The key difference is that, by adopting the algebraic approach, one may simultaneously consider all states arising in all Hilbert

space constructions of the theory without having to make a particular choice of representation at the outset. It is particularly important to proceed in this manner in, e.g., studies of phenomena in the early universe, so as not to prejudice in advance which states might be present.

The Microlocal Spectrum Condition Microlocal analysis provides a refined characterization of the singularities of a distribution by examining the decay properties of its Fourier transform. More precisely, let D be a distribution on a manifold, M , and let (x, k_a) be a point in the cotangent bundle of M . If D has the property that it can be multiplied by a smooth function, f , of compact support with $f(x) \neq 0$, such that the Fourier transform of fD decays more rapidly than any inverse power of $|k|$ in a neighborhood of the direction in Fourier transform space given by k_a , then D is said to be nonsingular at (x, k_a) . If D does not satisfy this property, then (x, k_a) is said to lie in the *wavefront set* (Hormander 1985), $\text{WF}(D)$, of D . In the case of quantum field theory in curved spacetime, the wavefront set can be used to characterize the singular behavior of the distributions $\omega[\phi_1(x_1) \dots \phi_n(x_n)]$ (as a subset of the cotangent bundle of $M \times \dots \times M$, where M is the spacetime manifold).

Now, for free fields in Minkowski spacetime, the positivity of total energy is directly related to the choice of positive frequency solutions in the decomposition (15.7). This, in turn, is directly related to the “locally positive frequency character” of the singular behavior of the n -point correlation functions $\omega[\phi(x_1) \dots \phi(x_n)]$ in the coincidence limit. Consequently, it can be shown that in Minkowski spacetime, the spectrum condition (positivity of total energy) is equivalent to a *microlocal spectrum condition* that restricts the wavefront set of $\omega[\phi_1(x_1) \dots \phi_n(x_n)]$. This microlocal spectrum condition can be generalized straightforwardly to curved spacetime. In this manner, it is possible to impose the requirement that states have a “locally positive frequency character” even in spacetimes where there is no natural global notion of “positive frequency” (i.e., no global notion of Fourier transform).

Local and Covariant Fields It is often said that in special relativity one has invariance under “special coordinate transformations” (i.e., Poincaré transformations), whereas in general relativity, one has invariance under “general coordinate transformations” (i.e., all diffeomorphisms). However, this is quite misleading. By explicitly incorporating the flat spacetime metric into the formulation of special relativity, it can easily be seen that special relativity can be formulated in as “generally covariant” a manner as general relativity, but the act of formulating special relativity in a generally covariant manner does not provide one with any additional symmetries or other useful conditions. The true meaning of “general covariance” is that the theory is constructed in a local manner from the spacetime metric and other dynamical fields, with no non-dynamical background structure (apart from manifold structure, and choices of space and time orientations and spin structure) playing any role in the formulation of the theory. This is the proper generalization of the notion of Poincaré invariance to general relativity.

In the present context, we wish to impose the requirement that in an arbitrarily small neighborhood of a point x , the quantum fields Φ under consideration “be locally and covariantly constructed out of the spacetime geometry” in that neighborhood. In order to formulate this requirement, it is essential that quantum

field theory in curved spacetime be formulated for *all* (globally hyperbolic) curved spacetimes—so that we can formulate the notion that “nothing happens” to the fields near x when we vary the metric in an arbitrary manner away from point x . The notion of a local and covariant field may then be formulated as follows (Brunetti et al. 2003): Suppose that we have a causality preserving isometric embedding $i : M \rightarrow \mathcal{O}' \subset M'$ of a spacetime (M, g_{ab}) into a region \mathcal{O}' , of a spacetime (M', g'_{ab}) . We require that this embedding induce a natural isomorphism of the quantum field algebra $\mathcal{A}(M)$ of the spacetime (M, g_{ab}) and the subalgebra of the quantum field algebra $\mathcal{A}(M')$ associated with region \mathcal{O}' . We further demand that under this isomorphism, each quantum field $\Phi(f)$ on M be taken into the corresponding quantum field $\Phi(i^* f)$ in \mathcal{O}' .

In what sense is this condition a replacement for Poincaré covariance? In the case of Minkowski spacetime, we can isometrically embed all of Minkowski spacetime into itself by a Poincaré transformation. The above condition then provides us with an action of the Poincaré group on the field algebra of Minkowski spacetime and also requires each quantum field in Minkowski spacetime to transform covariantly under Poincaré transformations. Thus, the above condition contains much of the essential content of Poincaré invariance, but it is applicable to arbitrary curved spacetimes without symmetries.

Let us now take stock of where things stand on the generalization of the basic principles of quantum field theory—as expressed by the Wightman axioms—to curved spacetime.

- The axiom that requires states to lie in a Hilbert space that carries a unitary representation of the Poincaré group is satisfactorily replaced by formulating theory via the algebraic approach and requiring that the quantum fields be local and covariant.
- The spectrum condition is satisfactorily replaced by the microlocal spectrum condition.
- The axiom stating that quantum fields are operator-valued distributions defined on a dense domain that is Poincaré invariant and invariant under the action of the fields and their adjoints is satisfactorily replaced by the GNS construction in the algebraic approach and the local and covariant field condition.
- The axiom that the fields transform in a covariant manner under the action of Poincaré transformations is satisfactorily replaced by the local and covariant field condition.
- As previously noted, the condition that at spacelike separations quantum fields either commute or anticommute generalizes straightforwardly to curved spacetime.

Thus, the only Wightman axiom that does not admit a satisfactory generalization to curved spacetime based on the ideas described above is the existence of a unique, Poincaré invariant state (“the vacuum”). This axiom plays a key role in the proofs of the PCT and spin-statistics theorem and many other results, so one would lose a great deal of the content of quantum field theory if one simply omitted this axiom. In particular, vacuum expectation values of products of fields play an important

role in many arguments, and it is crucial that these “c-number” quantities share the symmetries of the fields. On the other hand, we have already argued that it is hopeless to define a unique “preferred state” in generic spacetimes. Furthermore, states are inherently global in character and cannot share the “local and covariant” property of fields.

What is the appropriate replacement in curved spacetime of the requirement that there exist a Poincaré invariant state in Minkowski spacetime? Hollands and I have recently proposed (Hollands and Wald 2008) that the appropriate replacement is the existence of an operator product expansion of the quantum fields. An *operator product expansion* (OPE) is a short-distance asymptotic formula for products of quantum fields near point y in terms of quantum fields defined at y , i.e., formulae of the form

$$\phi^{(i_1)}(x_1) \cdots \phi^{(i_n)}(x_n) \sim \sum_{(j)} C_{(j)}^{(i_1) \cdots (i_n)}(x_1, \dots, x_n; y) \phi^{(j)}(y) \quad (15.12)$$

for all i_1, \dots, i_n , which hold as asymptotic relations as $\{x_1, \dots, x_n\} \rightarrow y$. The simplest example of an OPE is that for a product of two free Klein-Gordon fields in curved spacetime. We have

$$\phi(x_1)\phi(x_2) = H(x_1, x_2)\mathbf{1} + \phi^2(y) + \dots \quad (15.13)$$

where H is a locally and covariantly constructed Hadamard distribution (see, e.g., Kay and Wald (1991) or Wald (1994) for the precise form of H) and “...” has higher scaling degree than the other terms (i.e., it goes to zero more rapidly in the limit $x_1, x_2 \rightarrow y$). An OPE exists for free fields in curved spacetime and Hollands (2007) has shown that it holds order-by-order in perturbation theory for renormalizable interacting fields in curved spacetime. The requirement that an operator product expansion exists and satisfies a list of suitable properties (Hollands and Wald 2008) appears to provide an appropriate replacement for the requirement of the existence of a Poincaré invariant state. In particular, the distributional coefficients of the identity element in OPE expansions play much of the role played by “vacuum expectation values” in Minkowski spacetime quantum field theory.

By elevating the existence of an OPE to a fundamental status, Hollands and I have been led to the following viewpoint on quantum field theory in curved spacetime: The background structure, \mathcal{M} , of quantum field theory in curved spacetime is the spacetime (M, g_{ab}) , together with choices of time orientation, spacetime orientation, and spin structure. For each \mathcal{M} , we have an algebra $\mathcal{A}(\mathcal{M})$ of local field observables. In traditional algebraic approaches to quantum field theory, $\mathcal{A}(\mathcal{M})$ would contain the full information about the quantum field theory. However, in our approach, $\mathcal{A}(\mathcal{M})$, is essentially nothing more than the free $*$ -algebra generated by the list of quantum fields $\phi^{(i)}(f)$ (including “composite fields”), though it may be factored by relations that arise from the OPE.

In our viewpoint, all of the nontrivial information about the quantum field theory is contained in its OPE, Equation (15.12). The distributions $C_{(j)}^{(i_1) \cdots (i_n)}(x_1, \dots, x_n; y)$

appearing in Equation (15.12) are required to satisfy a list of properties, which include locality and covariance and an “associativity” property. States are positive linear maps on the algebra that satisfy the OPE relations as well as microlocal spectrum conditions.

Spin-statistics and PCT theorems have been proven within this framework (Hollands and Wald 2008). Interestingly, the PCT theorem relates processes in a given spacetime to processes (involving charge conjugate fields) in a spacetime with the opposite time orientation, e.g., it relates processes involving particles in an expanding universe to processes involving antiparticles in a contracting universe.

Renormalized perturbation theory can be carried out within this framework (Brunetti and Fredenhagen 2000; Hollands and Wald 2001, 2002). For a free field, composite fields (Wick powers)—such as ϕ^2 and T_{ab} —and all time-ordered products of fields can be defined in local and covariant manner. (However, “normal ordering” cannot be used to define composite fields.) The definition of time-ordered-products is unique up to “renormalization ambiguities” of the type expected from Minkowski spacetime analyses, but with additional local curvature ambiguities (which also occur in the definition of the composite fields). Theories that are renormalizable in Minkowski spacetime remain renormalizable in curved spacetime. Renormalization group flow can be defined in terms of the behavior of the quantum field theory under scaling of the spacetime metric, $g_{ab} \rightarrow \lambda^2 g_{ab}$ (Hollands and Wald 2003). Additional renormalization conditions can be imposed so that, in perturbation theory, the stress-energy tensor of the interacting field is conserved (for an arbitrary covariant interaction) (Hollands and Wald 2008).

In summary, the attempt to describe quantum field phenomena in curved spacetime has directly led to a viewpoint where symmetries and notions of “vacuum” and “particles” play no fundamental role. The theory is formulated in a local and covariant manner in terms of the quantum fields. This formulation is very well suited to investigation of quantum field effects in the early universe. In addition, the definition of composite fields, such as the stress-energy tensor, is intimately related to the OPE, and thus arises naturally in this framework. It is my hope that quantum field theory in curved spacetime will continue to provide us with deep insights into the nature of quantum phenomena in strong gravitational fields *and* into the nature of quantum field theory itself.

Acknowledgements This research was supported in part by NSF grant PHY04-56619 to the University of Chicago.

References

- Birrell, N. D., & Davies, P. C. W. (1982). *Quantum fields in curved space*. Cambridge: Cambridge University Press.
- Brunetti, R., & Fredenhagen, K. (2000). Microlocal analysis and interacting quantum field theories: renormalization on physical backgrounds, *Communications in Mathematical Physics*, 208, 623–661.

- Brunetti, R., Fredenhagen, K., & Verch, R. (2003). The generally covariant locality principle—a new paradigm for local quantum physics. *Communications in Mathematical Physics*, 237, 31–68.
- Hollands, S. (2007). The operator product expansion for perturbative quantum field theory in curved spacetime. *Communications in Mathematical Physics*, 273(1), 1–36.
- Hollands, S., & Wald, R. M. (2001). Local wick polynomials and time ordered products of quantum fields in curved spacetime. *Communications in Mathematical Physics*, 223(2), 289–326 .
- Hollands, S., & Wald, R. M. (2002). Existence of local covariant time ordered products of quantum fields in curved spacetime. *Communications in Mathematical Physics*, 231(2), 309–345.
- Hollands, S., & Wald, R. M. (2003). On the renormalization group in curved spacetime. *Communications in Mathematical Physics*, 237(1), 123–160.
- Hollands, S., & Wald, R. M. (2005). Conservation of the stress tensor in perturbative interacting quantum field theory in curved spacetime. *Reviews in Mathematical Physics*, 17, 227–312.
- Hollands, S., & Wald, R. M. (2008). Axiomatic quantum field theory in curved spacetime (2008). arXiv:0803.2003.
- Hormander, L. (1985). *The analysis of linear partial differential operators, I*. Berlin: Springer.
- Kay, B. S., & Wald, R. M. (1991). Theorems on the uniqueness and thermal properties of stationary, nonsingular, quasifree states on spacetimes with a bifurcate Killing horizon. *Physics Reports*, 207(2), 49–136.
- Streater, R. F., & Wightman, A. A. (1964). *PCT, spin and statistics and all that*. New York: Benjamin.
- Wald, R. M. (1994). *Quantum field theory on curved spacetimes and black hole thermodynamics*. Chicago: The University of Chicago Press.

Chapter 16

Conformal Infinity – Development and Applications



Jörg Frauendiener

16.1 Introduction

One of the major hallmarks of Einstein's theory of gravitation is the prediction of gravitational waves, see Einstein (1918). Einstein obtained this result very soon after he had published the final theory in Einstein (1916). As described in detail by Kennefick (1997), Einstein withdrew his claim of the existence of gravitational waves in the draft of a joint paper with Nathan Rosen because they had encountered singularities in the metric functions of a solution, which was supposed to describe plane waves. As it later turned out, the singularities were due to an unfortunate choice of coordinates and could be removed by using a different coordinate system.

In order to avoid such misconceptions it seemed appropriate to find characterisations of gravitational waves, which were independent of coordinates. The search for such an invariant way to describe gravitational waves was initiated only in the mid 1950s with the work of Pirani (1957) and continued for several years until 1963, the year in which a rigorous characterisation of radiative space-time was finally formulated. The ensuing geometric picture found numerous applications over the years and is still one of the most useful tools in the theory of gravity.

This article is dedicated to the memory of Jürgen Ehlers.

J. Frauendiener (✉)

Department of Mathematics and Statistics, University of Otago, P.O. Box 56, Dunedin 9010, New Zealand

Centre of Mathematics for Applications, University of Oslo, P.O. Box 1053, Blindern, NO-0316 Oslo, Norway

e-mail: joergf@maths.otago.ac.nz

© Springer Science+Business Media, LLC, part of Springer Nature 2018

D. E. Rowe et al. (eds.), *Beyond Einstein*, Einstein Studies 14,

https://doi.org/10.1007/978-1-4939-7708-6_16

In this contribution I will try to show how the discovery of the notion of conformal infinity came about from the first attempts to characterise radiation up to the full geometric picture. Furthermore, I will summarise what the current knowledge in this area of research is and present some of the applications. This work is a somewhat extended and simultaneously shortened version of Frauendiener (2004). Shortened, because the detailed treatment of the applications, in particular the numerical ones, will be suppressed – and extended because the development of the concept over time is discussed in more detail.

16.2 Towards the Invariant Characterisation of Gravitational Waves

As already mentioned in the introduction, there have been misunderstandings in the past as to what should constitute a gravitational wave. Einstein came to predict the existence of gravitational waves by analysing the field equations in a weak field approximation. He looked at linearisations of the field equations around flat space in a coordinate system, which was fixed by the condition that the field (in today's language) be transverse. With these assumptions he found a set of ten wave equations for the individual components γ_{ik} of the metric field, which he solved using the well-known integral representations of the fundamental retarded wave solutions.

Within the weak field approximation he determined the energy-momentum tensor of the gravitational field using the expressions for what is now called the energy-momentum pseudo-tensor. In general, this object is not well-defined, since it depends on the coordinate system (Frauendiener 1989). However, under the conditions that Einstein employed in his study the pseudo-tensor becomes well-defined, and it provides physically meaningful statements about the energy and momentum properties of the field configuration.

He then went on to discuss plane waves. Making the ansatz that the field depends only on the combination $x_1 + ix_4$ (i.e., $x \pm t$) he derived solutions of the linearised field equations, which contain six arbitrary constants. However, as he quickly points out, inserting these field configurations into the expressions for the energy-momentum tensor shows that of the six constants only two are physically relevant, because there are wave configurations which do not transport energy and must, therefore, be counted as unphysical. These unphysical 'modes' can be related to coordinate transformations. Einstein showed that for field configurations, which have no 'shear', i.e., for which $\gamma_{zz} - \gamma_{yy} = 0 = \gamma_{zy}$, one can find transformations such that the field components in the new coordinates vanish.

All this analysis has been carried out on the linearised level. Some years later, Beck (1925) derived the first exact solutions for plane and cylindrically symmetric waves. He discusses two kinds of plane waves, one of which is an unphysical

mode since it carries no energy, which he demonstrated by computing the energy-momentum pseudo-tensor. He also mentioned that these wave-like configurations can be transformed to flat space because the Riemann tensor vanishes identically in these situations. The other plane wave mode that he found does carry energy and he, therefore, counted it as physical. This argument has to be seen as a rigorous version of Einstein's previous discussion in a special highly symmetric case. Finally, Beck gives an exact solution for cylinder waves, but does not append a similar discussion about the physical nature of such waves.

Einstein and Rosen (1937) and Rosen (1937) take up the topic again. As described in illuminating detail by Kennefick (1997), they show with different arguments that plane waves cannot exist. The point of their proofs is that in such solutions singularities would necessarily appear. Furthermore, Taub (1951) and McVittie (1955) showed with different methods and assumptions that unpolarised plane wave solutions of the Einstein equations necessarily lead to flat space. For these reasons, and also due to the fact that the arguments involving the energy-momentum pseudo-tensor were not considered as conclusive, several authors, most notably Bondi and Scheidegger (1953), held the view that such solutions might not exist at all in the full (non-linear) theory, even though in the linear theory they do.

The new turn came with the works of Robinson, who showed that there was an error in Rosen's argument, followed by the papers Bondi (1957) and Bondi et al. (1959). They pointed out that the regularity conditions used by Einstein and Rosen were too restrictive, since they essentially assumed that the entire space-time manifold should be covered by a single regular coordinate chart. This point of view is of course not tenable, as already the simple case of a sphere in Euclidean space shows. The modern point of view, covering a manifold by several coordinate patches, had been used by Lichnérowicz (1958), who also gave less restrictive regularity conditions. Using these conditions Bondi and coworkers were able to construct physically reasonable plane wave configurations. They gave an argument that a plane 'sandwich' wave, i.e., a wave consisting of a non-flat region in an otherwise flat space-time bounded by null-hypersurfaces, will cause freely falling observers, initially at rest, to gain relative speeds after the impact of the wave. This argument shows that the wave carries energy, even though it was not clear at the time how to relate the intensity of the wave with the energy it carries.

Over time it became increasingly clear that issues involving the choice of coordinates were rather complicated and that it was desirable to have an invariant characterisation of radiation. The first attempts in this direction were undertaken by Pirani (1957). His point of view was to characterise radiation in terms of the Riemann tensor. The reason for this was based on the equivalence principle, which essentially implies that the physical effects of a gravitational field are due to its inhomogeneities, which in turn are measured by the Riemann tensor. This is very well illustrated by the argument described above of the effect of a sandwich wave on freely falling particles.

Pirani assumed that gravitational radiation in a vacuum propagates with the speed of light. This assumption was well backed by the result due to Lichnérowicz that the

characteristic surfaces of Einstein's equations are null-hypersurfaces. Pirani showed that discontinuities of the Riemann tensor propagate exactly along these characteristic surfaces and that the discontinuities are of the special algebraic types $\{211\}$, $\{31\}$ or $\{4\}$ (in modern language¹) i.e., at every point on the hypersurface the null-vector along the hypersurface points in a principal null direction of the Riemann tensor, which is at least twice degenerate. He also draws the analogy with electrodynamics, where the Faraday tensor for radiating systems can also be shown to have special algebraic properties. According to these results, Pirani defines the presence of gravitational radiation *at a space-time point* if and only if the Riemann tensor at that point is of those special algebraic types. In the discussion section of the paper, he notes that his definition applies at each point in space-time separately and that there is no implication about how the properties of the radiation changes as it propagates through space. He suggests that one should analyse this issue by using the Bianchi identities, which can be seen as a system of differential equations for the curvature linking the Riemann tensor at neighboring points. This is the first hint towards the importance of the Bianchi identities for understanding gravitational radiation. This suggestion was taken up by Sachs (1960), who investigated the propagation of algebraically special gravitational fields using the Bianchi identity for the Riemann tensor in vacuum regions. He also showed in a linear approximation that the far field of the Riemann tensor at large distances from a radiating source is a null-field, i.e., it has type $\{4\}$, while the 'semi-far' field has Petrov type III. This must be considered as a precursor to the full 'peeling property', which he derived only a year later (see below).

As Bondi (1957) noted, the plane waves he considered did in fact satisfy Pirani's definition. But already in Bondi et al. (1959) the authors mention that this criterion might be too restrictive. In discussions with C. Misner and A. Trautman it was found that Petrov's type II solutions, the type $\{4\}$ solutions, could only describe pure radiation, not the kind of radiation that could emanate from sources. They suggested that the definition should be weakened in the sense that one should require the Riemann tensor to be of Petrov type II only asymptotically. This means that one would expect the dominant terms of the Riemann tensor of a space-time including realistic material sources to be of type II at large distances from the sources.

This insight had been suggested by Trautman (1958a,b,c). He had been studying the question of boundary conditions for the Einstein equations, trying to obtain conditions that would allow for the presence of gravitational radiation from an isolated system, but still strong enough to guarantee uniqueness of solutions for reasonable data. This problem is not straightforward to address in general relativity, again because of the equivalence principle, which carries the implication that there is no background structure in the theory or, mathematically formulated, that the theory is invariant under arbitrary diffeomorphisms. In contrast to classical field theories,

¹Pirani uses Petrov's classification of the Riemann tensor in terms of three types, which are not enough to characterise all possibilities. Petrov type I corresponds to the types $\{1111\}$, $\{22\}$ and $\{-\}$, while Petrov type $\{II\}$ contains type $\{211\}$ and $\{4\}$ and Petrov's type II is type $\{31\}$.

which are built on geometrically fixed space-times, such as Minkowski space or a Newtonian space-time, in general relativity the geometry of space-time becomes a player in the game. This implies that there is no fixed ‘canvas’ that can be used to paint the field configurations. Instead, the canvas moves as well, and so one needs to invent certain devices that will ‘pin down’ the moving canvas.

In Trautman’s case, such a device was the introduction of a special class of coordinate systems with respect to which the metric approaches the flat Minkowski metric asymptotically (Trautman 1958b). Using these coordinates, he could formulate boundary conditions as fall-off conditions of the metric components, i.e., by specifying the rate at which they approach the Minkowski values in the asymptotic regime. He tried to mimic the situation in electrodynamics, where one knew that field configurations existed satisfying Sommerfeld’s *Ausstrahlungsbedingung*, the outgoing radiation condition. While it had been known at the time that solutions to the Maxwell equations subject to the Sommerfeld condition exist, this was unknown for the case of the Einstein equations.

In Trautman (1958c), Trautman went on to discuss the notion of energy in gravitation. Using Einstein’s pseudo-tensor for the energy-momentum of the gravitational field, he could show that – as a consequence of the boundary conditions on every space-like hypersurface which extended to infinity – there existed a well-defined energy-momentum 4-vector, which could be interpreted as giving the energy-momentum of the system at the instant defined by the hypersurface. He found that the difference between the expressions evaluated on different hypersurfaces was due to a flux-term through the time-like cylinder connecting the two at infinity. He concluded that this difference must be due to the radiation leaving the system towards infinity. Trautman (1958b) also investigated the relevance of Pirani’s definition for the presence of radiation at a point. As already mentioned above, he realised that this condition could only hold in the asymptotic regime, since this was the case for the fields satisfying his boundary condition.

Sachs (1961) takes up the matter where Trautman and Pirani left off. He notes that from their work it follows that the Riemann tensor of outgoing radiation fields at large distances from the source had approximately the same structure as the Riemann tensor of a plane wave, and that closer to the source there will be deviations from that algebraic structure and he aims towards analysing these deviations in order to find a covariant criterion for purely outgoing radiation. The most important steps in his analysis involve a detailed study of the geometry of outgoing null-geodesics, the introduction of an adapted null tetrad and, especially, the use of the Bianchi identity as a propagation equation for the curvature. These had already been used earlier in Sachs (1960) as well as by Jordan et al. (1961) for analysing the structure of special algebraic space-times. Sachs’ quite cumbersome analysis revealed the very peculiar behaviour of the Riemann tensor of space-times with geodesic rays, i.e., space-times that contain a congruence of null-geodesics. As one follows the affine parameter along these rays to infinity, the Riemann tensor exhibits what came to be called the *peeling property*.

The Riemann tensor of a vacuum space-time has the peeling property, if for any given null-geodesic with affine parameter r , which extends to infinity, the curvature

falls off along the curve as follows: to order $1/r$, the Riemann tensor is null (i.e. it has type $\{4\}$, which means it has a quadruple principal null direction (PND) tangent to the curve); to the next order $1/r^2$, it has type $\{3, 1\}$, with two PND's, one of them triple and aligned with the curve. For orders $1/r^3$ and $1/r^4$ the types are $\{211\}$ and $\{1111\}$, respectively. Thus, along the null-geodesic one can write the curvature symbolically as

$$C = \frac{\{4\}}{r} + \frac{\{31\}}{r^2} + \frac{\{211\}}{r^3} + \frac{\{1111\}}{r^4} + \mathcal{O}(r^{-5}).$$

The letter C on the left hand side of this equation denotes the Weyl tensor, i.e. the Riemann tensor of a vacuum space-time, and the terms in the numerators symbolise curvature tensors of the indicated type, which are independent of the affine parameter r . The $1/r^5$ -part is understood to be completely general in its algebraic structure and not related to the curve in any way. Thus, the PNDs 'peel off' the direction of the ray one by one, as one moves further and further into the interior of the space-time.

Sachs goes on to postulate that a space-time is free of incoming radiation if it contains asymptotically geodesic rays, which implies that its Riemann tensor obeys the peeling property. This is what he called the *outgoing radiation condition*. However, this is a condition that specifies the fall-off of the curvature at infinity by giving the structure of first terms in an expansion in the affine parameter. It is quite conceivable that this structure is not affected by incoming waves that remain well inside the space-time, such as an incoming radial sandwich wave or an incoming wave with a profile that dies off suitably fast towards infinity. The possible existence of incoming radiation of this kind has been found in later work by Bonnor and Rotenberg (1966) and by Newman and Penrose (1968).

A decisive step forward came with Bondi et al. (1962) and Sachs (1962b). These studies were concerned with an asymptotic analysis of the solutions of the field equations themselves. The main ingredient was a suitable ansatz for the metric. Previous work had made exceedingly clear that the null-cone structure of the space-time has a very tight relationship with the propagation of gravitational waves. So Bondi and his coworkers had the idea to introduce a coordinate system, which was adapted to this structure. The main idea was the definition of a retarded time coordinate u , whose level surfaces are null-hypersurfaces that open up towards the future and which regularly foliate the space-time away from the source regions. Assuming the existence of such a function u one can introduce an adapted so-called Bondi coordinate system. Each null-hypersurface given by constant values of retarded time is generated by null-geodesics. On each fixed hypersurface these null-geodesics can be labelled by two (angular) coordinates. As a fourth coordinate, Bondi et al. introduce a luminosity distance as radial coordinate along each null-generator. Furthermore, in order to simplify the calculations Bondi et al. (1962) assume symmetry of the field around a fixed axis, whereas Sachs (1962b) treats the general case.

Asymptotic conditions were imposed by requiring that one should be able to follow the null-geodesics towards infinity for arbitrarily large values of the radial coordinate and that the metric approaches the Minkowski metric in the limit of infinite r . As a further condition, it is required that the dependence of the metric components on $1/r$ is analytic. Another very useful property of this ansatz is the fact that the ensuing field equations have a hierarchical structure, which can successively be solved asymptotically near infinity. This process is essentially equivalent to the formulation and asymptotic solution of a certain characteristic initial value problem and the identification of its free data, as shown by Sachs in Sachs (1964). It turns out that there are two freely specifiable pieces of data, one complex-valued function of three variables at an arbitrarily chosen initial hypersurface $u = 0$ and another such function at “ $r = \infty$ ”. The latter function is Bondi’s *news function*, because it characterises the presence of gravitational radiation in a very precise sense.

The result of the asymptotic solution of the field equations turned out to be physically very reasonable. The main result was the demonstration that gravitational waves carry away energy-momentum from the system and hence diminish its mass. This is a consequence of the *Bondi-Sachs mass-loss formula*, which equates the rate of change of the mass at a retarded time u to the negative integral of the modulus of the news function over the boundary at infinity of the hypersurface of constant u . This latter integral is positive definite and vanishes if and only if the news function vanishes. It is interpreted as the integrated flux of gravitational radiation at infinity. A further consequence is that the Riemann tensor of the asymptotic solution does in fact satisfy Sachs’ peeling property and also Trautman’s asymptotic boundary condition. Finally, the group of coordinate transformations that leave the metric and the boundary conditions invariant was determined. This is an infinite-dimensional group, a semi-direct product of the Lorentz group with the infinite group of super-translations, nowadays known under the name Bondi-Metzner-Sachs (BMS) group, see Sachs (1962a). Its appearance came as a surprise because one might have expected the Poincaré group to appear instead of this strictly larger, albeit structurally very similar group. The BMS-group makes no reference to the specific metric used to derive it. Therefore, it must be regarded as the invariance group of a universal structure common to all space-times satisfying the Bondi-Sachs conditions at infinity.

At about the same time, when this study was underway in London, E. T. Newman and R. Penrose independently developed different formalisms for studying the equations of general relativity. Newman (1961a,b) introduced null-tetrads as basic variables into the theory; regarding the metric as a derived quantity, he formulated scalar equations for the components of the connection and the curvature in this basis. Penrose (1960), on the other hand, approached the topic from a different angle, making use of the well-known relationship between the Lorentz group and the group $SL(2, \mathbb{C})$. With this spinorial approach, he was able among other things to re-derive and to sharpen the Petrov classification of the Riemann tensor.

In their now classic paper (Newman and Penrose 1962), Newman and Penrose combined these two approaches into what is known today as the spin-coefficient or NP-formalism. They then apply this new formalism to the problem of gravitational

radiation. Using a coordinate system very similar to the one used by Bondi and Sachs, they solve the NP equations asymptotically for large values of r . They also show that the peeling-property follows from the single assumption that the Weyl tensor component $\Psi_0 = \mathcal{O}(r^{-5})$. This particular fall-off condition is consistent with the behaviour of the radiation from a time-dependent quadrupole in the linear theory. This result also improves those of Bondi and Sachs, since it no longer assumes analyticity of the metric coefficients in the inverse radius $1/r$. In a follow-up paper (Newman and Unti 1962), Newman and Unti determine even higher orders of the asymptotic solution of the field equation, thereby recovering the Bondi-Sachs mass-loss formula. In a certain sense, this paper by Newman and Unti marks the end of the period of more or less formal arguments. Some years later, Dixon (1970) showed that the algorithm developed by Newman, Penrose, and Unti can in principle produce the solution to all orders, so that there are no hidden obstructions to the solvability of the equation.

This period was followed by a phase in which the results found so far were geometrised and new concepts based on these insights were developed. The main contributions in this direction have undoubtedly been made by Roger Penrose. There were (at least) two major motivations for these developments: first, the analysis of the field equations had revealed that the curvature tensor exhibits a beautiful behaviour along null-directions toward infinity, thus indicating the importance of the location “ $r = \infty$ ”, whose structure, however, remained largely unexplored. Second, the importance of null-geodesics and null-hypersurfaces suggested that the conformal structure of space-time, i.e., its metric without a specific scale, played an important role. In fact, it had been stressed by E. Schücking that the zero-rest-mass equations with any spin are conformally invariant, and it was known from Penrose’s paper (Penrose 1960) that the Bianchi identity in a vacuum space-time is superficially identical to the spin-2 zero-rest-mass equation.

Penrose took up this challenge in Penrose (1963, 1964b, 1965). By studying space-times from the viewpoint point of their conformal structure, he develops a beautiful picture of the asymptotic properties of the space-time itself and of the fields defined on it. Ignoring the scale in the metric makes it possible to treat ‘infinity’ as a regular hypersurface bounding the space-time. Asymptotic behaviour is translated into local properties on the boundary hypersurface. Introducing an appropriate metric conformal to the original one, asymptotic calculations turn into local calculations on the boundary. Depending on the sign of the cosmological constant (+, −, 0), the hypersurface at infinity is time-like, space-like or null. In the null case, Penrose shows that the topology of the boundary hypersurface is the same as for Minkowski space-time. This conformal approach to asymptotic flatness turns out to be in remarkable agreement with the results of Bondi and Sachs, findings obtained by using a special class of coordinates. In particular, the peeling-property turns out to be a general feature of zero-rest-mass fields, thus a direct consequence of finiteness and continuity of the field on the boundary. Similarly, Bondi’s news function emerges as a certain geometric property of the boundary. The notion of gravitational radiation can now be uniquely defined as the value of the rescaled Weyl tensor on this boundary, i.e., a space-time is said to contain gravitational

waves if one can construct a conformal boundary on which this tensor field is non-zero. Finally, the BMS-group reappears in the conformal approach as the invariance group of the (degenerate) geometry of the boundary hypersurface.

Within this approach, a space-time is defined to be asymptotically flat if it has the same asymptotics as Minkowski space-time, i.e., if and only if it admits a conformal boundary with the same structure as for flat space-time. Apart from the obvious advantage of removing awkward asymptotic calculations in favour of well-defined local geometric considerations, the conformal approach opens up a completely new way to view asymptotics. At the very least, it provides some sort of universality in the sense that all asymptotically flat space-times have the same asymptotic structure. Or, in terms of the analogy used above, it provides a unique frame onto which the canvas of space-time can be mounted. In this way, it can be used to compare different space-times. It also exhibits a region of flatness in an otherwise curved space-time. This can be used under certain conditions to transfer constructions from flat space-times to asymptotically flat space-times. The prime example for this procedure is the concept of energy-momentum. It is well-known that there is no well-defined notion of gravitational energy-momentum density in a curved space-time. Yet, it turns out that one can define energy-momentum concepts uniquely for arbitrary asymptotically flat space-times.

Looking back at the development of the subject, it is striking how quickly things fell in place once the correct concepts were found. Apparently, it took people quite some time to understand that the important geometric structures were all related to the null-cone: null-hypersurfaces, null-geodesics, characteristic surfaces, etc. Once this was realised, it was only a matter of one or two years before a complete picture emerged. The decisive step came by realising the implications of the concept of null-infinity. Even in the papers by Sachs and Bondi et al., the locus “ $r = \infty$ ” is apparently implicitly conceived of as a time-like object. The precise structure of \mathcal{J} could only be correctly derived once the conformal rescaling method was adopted.

This development is remarkable also from the ‘sociological’ point of view. It would be quite interesting to follow the flow of ideas and information between the various groups at that time. There were only a handful of persons involved in the subject working together in various constellations and different venues. Without the modern methods of communication the information must have been transferred mostly by direct interaction. The atmosphere among the persons would probably have been quite intense, fuelled by the common desire to clear up the obscure nature of gravitational waves.

16.3 Asymptotic Structure and Conformal Geometry

In this section we will provide a short overview of the geometric features of conformal infinity. For a more detailed treatment we refer to the original papers by Penrose (1963, 1964a,b, 1965), the textbook (Penrose and Rindler 1986) and the reviews (Frauendiener 2004; Friedrich 1992, 1998a,b; Geroch 1977; Schmidt 1978).

16.3.1 Asymptotic Structure of Minkowski Space-Time

We start with Minkowski space-time (\mathbb{M}, η) with its flat metric η in polar coordinates

$$\eta = dt^2 - dr^2 - r^2 d\sigma^2, \quad (16.1)$$

where $d\sigma^2 = d\theta^2 + \sin^2\theta d\phi^2$ is the metric of the unit sphere S^2 . Introducing null coordinates $u = t - r$ and $v = t + r$ with $-\infty < u, v < \infty$ but subject to the restriction $v - u = 2r \geq 0$ the metric takes the form

$$\eta = dudv - r^2 d\sigma^2. \quad (16.2)$$

Now we introduce another set of null coordinates U and V , which have the purpose to map the infinite range, where u and v take their values to a finite range. There are several possibilities but it turns out that the definitions $u = \tan U$ and $v = \tan V$ work best. Now U and V are restricted by $-\pi/2 < U, V < \pi/2$ and $V \geq U$. The Minkowski metric takes the form

$$\eta = \frac{dUdV}{(\cos^2 U)(\cos^2 V)} - \frac{1}{4} \frac{\sin^2(V - U)d\sigma^2}{(\cos^2 U)(\cos^2 V)}. \quad (16.3)$$

Obviously, the metric is undefined at events where $\cos U = 0$ or $\cos V = 0$. These events do not belong to \mathbb{M} , since they would have $u = \pm\infty$ or $v = \pm\infty$. However, defining the function

$$\Omega = 2 \cos U \cos V \quad (16.4)$$

on \mathbb{M} we see that the metric

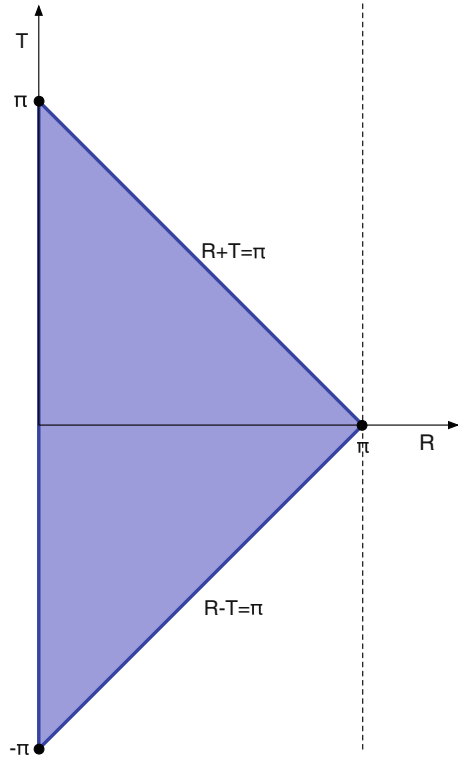
$$\hat{\eta} = \Omega^2 \eta = 4dUdV - \sin^2(V - U) d\sigma^2 \quad (16.5)$$

is regular for all values of V and U with $V \geq U$. The metric $\hat{\eta}$ is conformally equivalent to η . A final coordinate transformation re-introduces a time coordinate $T = U + V$ and a radial coordinate $R = V - U$, which are restricted by the conditions $-\pi < T < \pi$ and $R \geq 0$. In these coordinates $\hat{\eta}$ takes the form

$$\hat{\eta} = dT^2 - dR^2 - \sin^2 R d\sigma^2, \quad (16.6)$$

which turns out to be the metric of Einstein's static universe **E**. Figure 16.1 shows the 2-dimensional (T, R) -plane obtained by ignoring the angular dependence of the metric. The points between the lines $R = 0$ and $R = \pi$ correspond to events in **E**, while points inside the shaded region correspond to points in \mathbb{M} . This region is

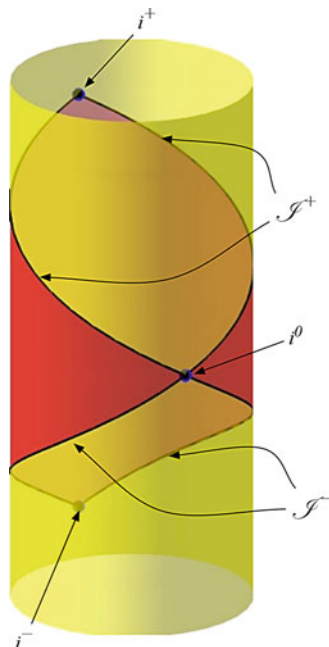
Fig. 16.1 The (T, R) -plane



bounded by the lines $2V = T + R = \pi$, $2U = T - R = -\pi$ and $U = V$. Points on the line $U = V$ have $R = 0$ and also $r = 0$, i.e., they correspond to events on the time-like world-line of the origin for the polar coordinates in Minkowski space-time.

Strictly speaking, most points in the diagram represent a 2-sphere $(T, R) = \text{const}$, except for the points with $R = 0$ and $R = \pi$ because at these points the radius of the corresponding 2-sphere shrinks to zero. This diagram tells us that we may regard \mathbb{M} as being embedded into the larger Einstein cylinder \mathbb{E} in this very particular way. Points on the line $U = -\pi/2$ have $u = -\infty$, while points on $V = \pi/2$ have $v = \infty$. Thus, they do not belong to \mathbb{M} , but they are clearly at the boundary $\partial\mathbb{M}$ of \mathbb{M} , when regarded as a part of \mathbb{E} . Note, that the boundary consists of five different parts. The (open) lines $|T| + R = \pi$ correspond to a 3-dimensional null-hypersurface each, while the end points $(0, \pm\pi)$ and $(\pi, 0)$ are points. A somewhat better rendering of the geometric situation is given in Figure 16.2. Although \mathbb{M} is diffeomorphic to the shaded part of the cylinder in Figure 16.2, they are not isometric. This follows easily when one considers the behaviour of geodesics. Fix a point P inside \mathbb{M} and follow a null-geodesic with respect to $\hat{\eta}$ through P towards the future. It will intersect the boundary $\partial\mathbb{M}$ after a finite amount of its affine parameter has elapsed. However, if this is done with a null-

Fig. 16.2 The embedding of Minkowski space-time into the Einstein cylinder



geodesic with respect to the Minkowski metric η through P in the same direction then the curve will never reach the boundary for any amount of affine parameter. Thus, the boundary lies “at infinity” with respect to η but at a finite distance from P with respect to $\hat{\eta}$.

When we consider η -geodesics from all possible points emanating in all possible directions we find that there are five different cases corresponding to the five different pieces of the boundary $\partial\mathbb{M}$. All space-like geodesics come close to the point i^0 when their affine parameter extends arbitrarily towards $\pm\infty$. Time-like geodesics approach the point i^\pm , when their affine parameter tends to $\pm\infty$. For this reason, the points i^+ and i^- are called *future* and *past time-like infinity*. Finally, each future/past directed null-geodesic approaches one single point on the hypersurface $\mathcal{J}^+/\mathcal{J}^-$, which is, therefore, called *future/past null-infinity*.² The two parts of null-infinity are regular null-hypersurfaces in \mathbf{E} (they are given by an equation $U = -\pi/2$ or $V = \pi/2$), while the three points i^\pm and i^0 are regular points of \mathbf{E} in the sense that the metric (16.6) is well-defined there. This is not automatic, since at these points infinitely many geodesics are squashed into one point. However, the flatness of \mathbb{M} guarantees that the geodesics approach at exactly the right rate for the points to be regular. A closer look reveals that the null-geodesics, which generate \mathcal{J}^- emerge from i^- , thus forming a regular null-cone, and reconverge in

²The letter \mathcal{J} is a ‘ \mathcal{S} ’, from which the colloquial name ‘scri’ for null-infinity derives.

i^0 . Similarly, for \mathcal{J}^+ , which is generated by null-geodesics emerging from i^0 and converging in i^+ .

This example shows that the geometry of the boundary is determined by the geometry of the metric η of \mathbb{M} . Had we chosen a different function $\tilde{\Omega} = \omega\Omega$ with some strictly positive function ω , then we would have obtained the exact same structure of the boundary $\partial\mathbb{M}$. However, beyond the boundary the geometry might have been different, i.e., we would not have necessarily obtained the Einstein cylinder but some different Lorentzian manifold.

16.3.2 Asymptotically Flat Space-Times

The example of the asymptotic structure of Minkowski space-time motivates the definition of an asymptotically flat space-time. The idea is simply that a space-time should be called asymptotically flat, if it admits a conformal boundary, which has the same properties as the one for flat space. This leads to the following

Definition 1 A space-time $(\tilde{\mathcal{M}}, \tilde{g})$ is called asymptotically flat if and only if there exists a manifold \mathcal{M} with smooth boundary $\mathcal{J} = \partial\mathcal{M}$, metric g and scalar field Ω , such that

- (i) $\tilde{\mathcal{M}}$ is diffeomorphic to the interior $\text{int}\tilde{\mathcal{M}} = \mathcal{M} - \mathcal{J}$ of \mathcal{M} ,
- (ii) $g_{ab} = \Omega^2 \tilde{g}_{ab}$,
- (iii) $\Omega > 0$ on $\tilde{\mathcal{M}}$ and $\Omega = 0, d\Omega \neq 0$ on \mathcal{J} ,
- (iv) the Ricci tensor \tilde{R}_{ab} of \tilde{g}_{ab} vanishes in a neighbourhood of \mathcal{J} .

There are some more technical requirements needed to ensure that this definition is not empty and to include also the physically interesting cases of black-hole space-times, see Penrose and Rindler (1986). Condition (i) states that we should regard $\tilde{\mathcal{M}}$ as part of \mathcal{M} , or, equivalently, that \mathcal{M} is obtained by adding boundary points to $\tilde{\mathcal{M}}$. Condition (ii) establishes the conformal equivalence of the metrics g_{ab} and \tilde{g}_{ab} on \mathcal{M} , while (iii) ensures that the boundary \mathcal{J} is a regular hypersurface of \mathcal{M} . Finally, from condition (iv) it follows in view of the Einstein equations that the asymptotic region is matter free. Again, this condition could be weakened by requiring suitable fall-off conditions for the energy-momentum tensor in $\tilde{\mathcal{M}}$.

Let us now look at the geometry of \mathcal{J} in more detail. In order to do this we need to relate the curvature tensors of the two metrics. The relevant formulae are

$$\tilde{\Lambda} = \Omega^2 \Lambda - \frac{1}{4}\Omega\Box\Omega + \frac{1}{2}\nabla_a\Omega\nabla^a\Omega, \tag{16.7}$$

$$\tilde{\Phi}_{ab} = \Phi_{ab} + \frac{1}{\Omega} \left(\nabla_a\nabla_b\Omega - \frac{1}{4}g_{ab}\Box\Omega \right), \tag{16.8}$$

$$\tilde{C}^a{}_{bcd} = C^a{}_{bcd}, \tag{16.9}$$

relating (multiples of) the scalar curvatures Λ and $\tilde{\Lambda}$, the trace-free parts Φ_{ab} and $\tilde{\Phi}_{ab}$ of the Ricci tensors as well as the Weyl tensors. Here, ∇_a denotes the covariant derivative with respect to the metric g_{ab} and $\square = \nabla^a \nabla_a$ is the corresponding wave-operator. Focussing first on (16.7) it follows from condition (iv) that on \mathcal{J} , where $\Omega = 0$,

$$\nabla_a \Omega \nabla^a \Omega = 0$$

holds. This implies that the hypersurface \mathcal{J} given by $\Omega = 0$ is a *null* hypersurface. It follows, that the normal vector field $N^a = -g^{ab} \nabla_b \Omega$ is also tangent to \mathcal{J} . As a null-hypersurface it is generated by a twist-free congruence of null-geodesics with tangent vector N^a . Further information about the properties of this hypersurface are obtained from (16.8). Using again the condition that $\tilde{\mathcal{M}}$ is vacuum near \mathcal{J} one has

$$0 = \Omega \Phi_{ab} + \nabla_a \nabla_b \Omega - \frac{1}{4} g_{ab} \square \Omega,$$

which implies that on \mathcal{J}

$$\nabla_a N_b = \frac{1}{4} g_{ab} \nabla^c N_c,$$

i.e., $\nabla_a N_b$ is symmetric and pure trace. Therefore, the null-hypersurface \mathcal{J} is *shear-free*. It is a little bit more involved to show that one can always choose a conformal factor Ω in such a way that the congruence of generators of \mathcal{J} is also *divergence-free*.

Following Sachs (1961), one can regard a null-congruence as a family of light rays. Shear and divergence of the congruence determine the shape of the image of an object mapped by these rays: a divergent congruence magnifies the image, while a shearing congruence causes astigmatism. For \mathcal{J} this means that an object is mapped by its generators without any distortion to its image. The lack of astigmatism is an intrinsic property, while the magnification properties of \mathcal{J} are related to the choice of the conformal factor. This is in line with Sachs' result that the presence of a gravitational field creates shear and divergence in a null-congruence. The lack of those on \mathcal{J} implies that there is no gravitational field present.

An important consequence of Definition 1, which is more difficult to obtain, and which depends in an essential way on the fact that \mathcal{J} has topology $S^2 \times \mathbb{R}$ is

Proposition 1 *The Weyl tensor $C^a{}_{bcd}$ vanishes on \mathcal{J} .*

This may be regarded as the ultimate justification for the term ‘asymptotically flat’: assuming that the energy-momentum tensor (and hence the Ricci tensor) vanishes (or falls off sufficiently fast) at infinity implies that the entire Riemann tensor vanishes on \mathcal{J} — space-time is flat at infinity.

The Weyl tensor of the metric \tilde{g}_{ab} plays a crucial role: on $\tilde{\mathcal{M}}$ it satisfies the *Bianchi equation*

$$\tilde{\nabla}_a \tilde{C}^a{}_{bcd} = 0. \quad (16.10)$$

a consequence of the Bianchi identity and the vacuum field equation, and it agrees there with the Weyl tensor $C^a{}_{bcd}$ of the metric g_{ab} due to the conformal invariance (16.9). Due to Prop. 1 the tensor field

$$K^a{}_{bcd} = \Omega^{-1} C^a{}_{bcd}$$

is well-defined on \mathcal{M} and as a consequence of the conformal transformation of the covariant derivative operator it satisfies the equation

$$\nabla_a K^a{}_{bcd} = 0. \quad (16.11)$$

This looks like the zero-rest-mass equation for a spin-2 field on \mathcal{M} . In fact, this tensor field most directly describes the gravitational field. In particular, its values on \mathcal{J} are closely related to the gravitational radiation escaping to infinity. From (16.11) and the regularity of \mathcal{J} follows that the field $K^a{}_{bcd}$ has a very particular behaviour near \mathcal{J} , which, when translated back to the Weyl tensor on \mathcal{M} , corresponds exactly to Sachs' peeling-property.

It is remarkable how Einstein's theory manages to create a frame at infinity for the 'canvas' of space-time. The lowest order in Ω serves to set up the conformal boundary with its basic flat structure on which the Weyl tensor vanishes. This structure is universal for all asymptotically flat space-times. The field $K^a{}_{bcd}$, which is the next order in Ω in the Weyl tensor, propagates on that background. One can go much further in the analysis of the structure of \mathcal{J} and the fields induced on it. The main tool is the Bianchi equation (16.11), which can be evaluated on \mathcal{J} in increasing orders assuming that the field is sufficiently smooth.

16.4 Existence of Asymptotically Flat Space-Times

It is clear that the method of conformal compactification provides a useful and unified framework for the discussion of asymptotic properties and, in particular, gravitational radiation. However, what is not so clear is whether there are indeed non-trivial space-times which admit a conformal boundary according to Penrose's procedure. After all, there are two different requirements that such space-times need to satisfy. On the one hand, they need to have a very specific geometric property, namely the provision for attaching the conformal boundary, which translates into very specific fall-off properties of the metric and its derivatives. On the other hand, they must satisfy Einstein's field equation, a partial differential equation for the metric. Thus, the question arises as to what extent these two conditions are compatible.

One can show directly, that apart from Minkowski space-time both the Schwarzschild and Kerr space-times admit a conformal boundary. In contrast to flat space in these space-times the points i^\pm and i^0 are singular. More generally, all stationary space-times have a smooth \mathcal{J} . However, none of these cases exhibit gravitational radiation. Radiating examples include the boost-rotation symmetric space-times, which have been explored in detail by Bičák and Schmidt (1989). In this case, however, the conformal boundary is not complete. Cutler and Wald (1989) demonstrated the existence of radiating solutions within Einstein-Maxwell theory, which are smooth on both null-infinities as well as future and past time-like infinity.

For a long time it was not clear whether there are any generic radiative asymptotically flat space-times in the sense of Definition 1, or whether the examples exist only due to special circumstances. To answer this question, two different avenues have been pursued. For the sake of brevity, let us call them the ‘analytical’ and the ‘geometrical’ approaches. The ‘analytical’ way deals with the Einstein equations in the physical non-rescaled space-time. Christodoulou and Klainermann (1993) proved the non-linear stability of Minkowski space. This means that ‘close’ to flat space-time there are non-trivial vacuum space-times, i.e., solutions of the vacuum Einstein equation, which have the same global structure as Minkowski space-time. In particular, these space-times have null-infinities but not, in general, regular space- and time-like infinities. In the first version of this result, there had been some discrepancies between the fall-off rates for the Weyl tensor and what was expected from the peeling property, i.e., it could not be guaranteed that \mathcal{J} was in fact smooth enough to allow for the peeling property. However, this early result has been improved upon by several authors using different methods, and nowadays there is no doubt that asymptotically flat space-times in the sense of Definition 1 exist, see Klainermann and Nicolò (2003); Lindblad and Rodnianski (2005) and the references therein.

The ‘geometrical’ way is more in line with Penrose’s original geometric approach; (for a more detailed review we refer to Friedrich 2004). Instead of starting with the physical space-time $\tilde{\mathcal{M}}$, one views the problem in the rescaled, ‘unphysical’ space-time \mathcal{M} . To do this, it is of course necessary to have field equations for the geometry of \mathcal{M} , which impose the condition that on $\tilde{\mathcal{M}}$ the vacuum equations are satisfied. This is not quite straightforward, since a look at (16.7), (16.8) shows that the unphysical Ricci tensor will contain terms that are formally singular on \mathcal{J} . Hence, a different point of view is needed. Friedrich (1979) presents a first order system of equations for the conformal metric g_{ab} with conformal factor Ω and Levi-Civita connection and curvature on the conformal manifold \mathcal{M} . These equations are regular on \mathcal{M} , in particular on the locus $\Omega = 0$. The system consists of various groups of equations: the compatibility condition between the metric and the connection, the equations relating the curvature to the derivatives of the connection coefficients, the Equation (16.8) regarded as a differential equation for Ω and the Bianchi equation (16.11). If a metric g_{ab} , together with its connection and curvature and a scalar field Ω satisfy this system of equations, then the metric $\tilde{g}_{ab} = \Omega^{-2}g_{ab}$, defined wherever $\Omega \neq 0$, will be a vacuum metric. As is to be

expected, the system is invariant under the rescaling $g_{ab} \mapsto \Theta^2 g_{ab}$, $\Omega \mapsto \Theta \Omega$ for any positive function Θ .

There exist several versions for these so-called *conformal field equations*. They can be formulated in terms of the metric, or in terms of an orthonormal tetrad as a basic variable, just as this is done for the standard Einstein equations. A more drastic step is to free up the connection by relaxing the requirement that this be the Levi-Civita connection for the metric (Friedrich 1998c). Indeed, the physical metric \tilde{g}_{ab} defines a conformal class of metrics, of which the unphysical metric g_{ab} is just one representative member. Related to this conformal class there is a class of Weyl connections, differing from each other essentially by a 1-form, of which the Levi-Civita connections of the metrics are equivalent representatives. By allowing for an arbitrary Weyl connection in the formulation of the conformal field equations, one gains additional freedom that can be used to advantage in certain situations, see Friedrich (1995, 2002, 2003).

The conformal field equations can be treated like any other system of geometrical partial differential equations. They split into constraints and evolution equations, which are compatible in the sense that the evolution equations propagate solutions to the constraint equations. They contain gauge freedom, which allows one to choose coordinates and the conformal gauge function Θ and, if applicable, a frame field and a Weyl connection. All these quantities have to be fixed before one can make statements about the uniqueness of solutions. Once they are fixed, the evolution equations can be put into a symmetric hyperbolic system. Then, using standard theory one has existence and uniqueness of solutions of the evolution equations local in time. The appropriate Cauchy problem in this context is the so-called *hyperboloidal initial-value-problem*, where initial data (satisfying the constraint equations) are prescribed on a space-like hypersurface Σ in \mathcal{M} , which intersects \mathcal{J}^+ transversely in a space-like 2-sphere.³ Friedrich (1986) shows that initial data sufficiently close to flat initial data develop into an asymptotically flat space-time for which null-infinity forms a regular cone with a vertex i^+ representing time-like infinity. Furthermore, Andersson, Chruściel, and Friedrich (Andersson et al. 1992) demonstrate that the constraint equations on hyperboloidal hypersurfaces can be solved so that enough initial data can always be found for the conformal evolution equations.

These results give semi-global existence; they guarantee completeness in the future (or, alternatively, in the past). There is one further step necessary in order to prove global existence of solutions: one needs to show that initial data prescribed on an asymptotically flat hypersurface evolve far enough into the future and into the past to contain hyperboloidal hypersurfaces intersecting \mathcal{J}^+ resp. \mathcal{J}^- . This can then be used to pose the hyperboloidal initial-value-problem. Chruściel, Piotr T. and

³The name arises from the fact that when the part of Σ lying inside $\tilde{\mathcal{M}}$ is considered as a hypersurface with induced geometry from the physical metric \tilde{g}_{ab} one finds that it is a manifold with asymptotically constant negative scalar curvature, just like the space-like hyperboloids in Minkowski space.

Delay achieved this in Chruściel and Delay (2002), building on a result obtained by Corvino (2000), who developed a new approach to solving the constraint equations on asymptotically flat hypersurfaces. For further details we refer to Friedrich (2004).

The existence of non-trivial asymptotically flat space-times which incorporate Penrose's proposal of conformal infinity has been established with both approaches mentioned. While there are still some open questions concerning the exact behaviour of space-times near space-like infinity i^0 (see Friedrich 1998c), this does not prohibit the use of the notion of conformal infinity in several applications.

16.5 Applications

16.5.1 *Isolated Systems*

The most important application of null-infinity springs from its roots, gravitational radiation. The current efforts to directly identify gravitational waves are based on a network of gravitational wave detectors (mostly of the interferometric type) waiting to catch a signal from various possible sources. The theoretical model which underlies this experiment and the interpretation of the data rests upon a very important idealisation, the *isolated system* (Geroch 1977). This refers to a self-gravitating system, such as a binary system, or a super nova etc., under observation e.g., by gravitational wave detectors. In reality, the system is in interaction with all parts of the universe, including the observers, so that, strictly speaking, there is no clear cut way to say where the system ends and the rest of the universe begins. However, in many interesting circumstances some parts will interact much more strongly with each other than with other parts and one would like to focus attention on these. This means that one ignores all the interactions with the rest and in particular with the observers. One assumes, in effect, that the system is 'alone in the universe'. Since the interactions are gravitational, this means that the gravitational field will decay with increasing distance from the system so that the space-time in the end will become flat. Thus, the gravitational wave experiment is modeled on an asymptotically flat space-time. The source generates the curvature of the space-time and the observers are assumed to be 'at infinity'. In fact, the observers correspond to the generators of \mathcal{J}^+ (Frauendiener 2000).

The idealisation of a radiating system as an isolated system, and hence as an asymptotically flat space-time, underlies in one form or another all attempts to compute the gravitational wave signals emitted from astrophysical sources.⁴ It is built into approximation schemes such as the post-Newtonian approximations. It is also used to establish boundary conditions for numerical codes, and it is used

⁴Strictly, speaking this refers only to non-stochastic sources.

to devise various ‘wave extraction schemes’ to isolate the radiative information from the computed data. All this is possible because an asymptotically flat space-time is characterised by the existence of null-infinity, and because there, and only there, the radiation is unambiguously defined. Computations based on the standard Einstein equations, i.e., those, which make no use of the possibility of conformal rescaling, thereby necessarily have to make a further approximation because they need to approximate the infinitely distant null-infinity by some artificial boundary. Computations based on the conformal field equations, however, do not have this problem. They can handle the boundary at infinity in terms of the conformal metric g_{ab} , for which null-infinity corresponds to a finite location, cf. Figure 16.2. Unfortunately, this approach has not yet been explored sufficiently, see Frauendiener (2004) for further discussions.

16.5.2 Other Applications

Since its discovery, null-infinity has played a major role in the development of theoretical concepts in general relativity. The reason for its importance is due to the fundamental problem of the theory of gravity: there is no fixed background on which the dynamics of the gravitational field evolves. The geometry itself is dynamic. Only on \mathcal{J} is there some sort of ‘universal flatness’ which can be used as a substitute for the flat Minkowski space-time.

One line of research starts out with the observation that in Minkowski space each event $P \in \mathbb{M}$ leaves a trace on \mathcal{J}^+ : consider the light-cone from P and follow it out to infinity. The light-cone will ultimately intersect \mathcal{J} in a 2-dimensional space-like surface which is topologically a sphere. For each event there is a different so-called ‘cut’ of \mathcal{J} . So the question arises: how can one characterise those cuts that correspond to events and, conversely, can one use such a characterisation to reconstruct the events in an arbitrary asymptotically flat space-time? This was an important question in connection with the quantisation of space-times.

It turns out that the cuts corresponding to space-time events in \mathbb{M} are shear-free, i.e., the congruence of null-geodesics generating the light-cones has no shear. The condition for a cut to be shear-free (a ‘good’ cut) turns out to be a non-linear second-order equation on a sphere, the so-called ‘good cut equation’. In the case of a flat space, this equation has a 4-parameter family of solutions, and these can be identified with the space-time events whose light-cones intersect null-infinity in the corresponding cut. In the general case, the good cut equation no longer has a 4-parameter family of solutions. However, Newman (1976) found the remarkable result that by complexifying the equation one obtains a 4-dimensional solution space that can be made into a complex manifold. Furthermore, this space carries a naturally defined metric for the distance between two complex cuts of complexified null-infinity, and even more remarkable, this metric is Ricci flat. Thus, in the general case, the good cut equation leads to a virtual or ideal vacuum space-time, whose

points are the good cuts and which is, in general, complex. Only in the flat case, will this ideal space-time – aptly called ‘heaven’ or \mathcal{H} -space – coincide with a real space-time.

The structure of the space of cuts of \mathcal{J} is still being actively explored with some surprising consequences, such as certain properties of complex world-lines found by Newman and his colleagues. These form 1-parameter solutions of certain equations for cuts on null-infinity in a space-time satisfying the Einstein-Maxwell equations. They can thus be interpreted as centres of mass and charge. They also obey equations of motion, which automatically contain, remarkably, the same radiation reaction forces that one obtains from classical electrodynamics for a charged particle, see Newman et al. (2008).

Motivated in part by the fundamental importance of the conformal structure of space-times, Penrose developed *twistor theory* as an alternative approach towards the quantisation of gravity, see Penrose (1999) for a review. In twistor theory, space-time structures are not considered as being fundamental. Instead they are derived from the more fundamental concept of a twistor. Very simplistically, a twistor can be thought of as a light-ray and a space-time point, the latter being the intersection of at least two light-rays, i.e., twistors. In this way, one obtains a non-local correspondence between space-time and the space of twistors. In the case of a flat space-time there exists a very powerful transformation by which structures on Minkowski space-time can be transformed to twistor space. This twistor correspondence has led to remarkable consequences in several areas of mathematics, see the contributions in Huggett et al. (1998). Unfortunately, so far all attempts to establish such a correspondence for general space-times have failed. However, for asymptotically flat space-times, the flatness of null-infinity again comes to the rescue, since one can construct an asymptotic twistor space; there is a tentative proposal for deriving space-time points from these asymptotic twistors.

A final example of the use of null-infinity can be found in Ashtekar’s asymptotic quantisation. Here, one takes advantage of the universality of \mathcal{J} as a background on which fields, in particular the gravitational field, propagate. Using \mathcal{J} as the kinematical arena, one can set up a phase space for the radiative modes of the gravitational field, i.e., equip the space of fields with a symplectic structure. One can also define ‘positive frequency’ fields on \mathcal{J} by means of which a Hilbert space of radiative modes is constructed, see Ashtekar (1987). Canonical commutation relations are imposed on the news function, and one can construct the Fock space in the usual way. Other representations of the CCR exist, which are based on the so-called *memory* (Christodoulou 1991; Frauendiener 1992).

16.6 Conclusion

Penrose’s notion of conformal infinity has brought a completely new insight into the geometry of the asymptotic regions of an asymptotically flat space-time. It has changed the way we look at such space-times today and it forms the foundation

for many developments, such as Newman's \mathfrak{H} -space, Penrose's twistor theory and Ashtekar's asymptotic quantisation procedure. It arose in the attempt to rigorously understand the nature of gravitational waves. The early characterisations of the class of radiative space-times in terms of boundary conditions and asymptotic fall-off conditions led to a beautiful and useful geometric concept of an asymptotically flat space-time. The Einstein equations are compatible with these geometric ideas and many examples for radiative space-times exist. Null-infinity has played a major role in the development of several theoretical concepts in gravitation and it continues to do so nowadays in applied areas such as the computation of gravitational wave signals from isolated sources.

Acknowledgements The author is very grateful to the organisers of the conference 'Beyond Einstein' for the opportunity to present the material of this contribution. Furthermore, it is a pleasure to thank Ted Newman, Roger Penrose and Engelbert Schücking for sharing their memories of the development of the subject.

References

- Andersson, L., Chruściel, P. T., & Friedrich, H. (1992). On the regularity of solutions to the Yamabe equation and the existence of smooth hyperboloidal initial data for Einstein's field equations. *Communications in Mathematical Physics*, *149*, 587–612.
- Ashtekar, A. (1987). *Asymptotic quantization*. Naples: Bibliopolis.
- Beck, G. (1925). Zur Theorie binärer Gravitationsfelder. *Zeitschrift für Physik*, *33*, 713–728.
- Bičák, J., & Schmidt, B. G. (1989). Asymptotically flat radiative space-times with boost-rotation symmetry. *Physical Review D: Particles and Fields*, *40*, 1827–1853.
- Bondi, H. (1957). Plane gravitational waves in general relativity. *Nature*, *179*, 1072–1073.
- Bondi, H., Pirani, F. A. E., & Robinson, I. (1959). Gravitational waves in general relativity III. Exact plane waves. *Proceedings of the Royal Society of London A*, *251*, 519–533.
- Bondi, H., van der Burg, M. G. J., & Metzner, A. W. K. (1962). Gravitational waves in general relativity VII. Waves from axi-symmetric isolated systems. *Proceedings of the Royal Society of London A*, *269*, 21–52.
- Bonnor, W. B., & Rotenberg, M. A. (1966). Gravitational waves from isolated sources. *Proceedings of the Royal Society of London A*, *289*, 247–274.
- Christodoulou, D. (1991). Nonlinear nature of gravitation and gravitational-wave experiments. *Physical Review Letters*, *67*, 1486–1489.
- Christodoulou, D., & Klainermann, S. (1993). *The global nonlinear stability of the Minkowski space*. Princeton: Princeton University Press.
- Chruściel, P. T., & Delay, E. (2002). Existence of non-trivial, vacuum, asymptotically simple space-times. *Classical and Quantum Gravity*, *19*, L71–L79.
- Corvino, J. (2000). Scalar curvature deformation and a gluing construction for the Einstein constraint equations. *Communications in Mathematical Physics*, *214*, 137–189.
- Cutler, C., & Wald, R. M. (1989). Existence of radiating Einstein-Maxwell solutions which are c^∞ on all of J^- and J^+ . *Classical and Quantum Gravity*, *6*, 453–466.
- Dixon, W. G. (1970). Analysis of the Newman-Unti integration procedure for asymptotically flat space-times. *Journal of Mathematical Physics*, *11*, 1238–1248.
- Einstein, A. (1916). Die Grundlage der allgemeinen Relativitätstheorie. *Annalen der Physik*, *49*, 769–822.

- Einstein, A. (1918). *Über Gravitationswellen* (pp. 154–167). Berlin: Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften.
- Einstein, A., & Rosen, N. (1937). On gravitational waves. *Journal of the Franklin Institute*, 223, 43–54.
- Frauendiener, J. (1989). Geometric description of energy-momentum pseudotensors. *Classical Quantum Gravity*, 6, L237–L241.
- Frauendiener, J. (1992). Note on the memory effect. *Classical Quantum Gravity*, 9, 1639–1641.
- Frauendiener, J. (2000). Numerical treatment of the hyperboloidal initial value problem for the vacuum Einstein equations. III. On the determination of radiation. *Classical Quantum Gravity*, 17, 373–387.
- Frauendiener, J. (2004). Conformal infinity. *Living Reviews in Relativity*, 7, 1. <http://www.livingreviews.org/lrr-2004-1>.
- Friedrich, H. (1979). On the regular and the asymptotic characteristic initial value problem for Einstein's vacuum field equations. In M. Walker (Ed.), *Proceedings of the third Gregynog relativity workshop, gravitational radiation theory*. Max-Planck Green Report, MPI-PAE/Astro 204, Oktober 1979.
- Friedrich, H. (1986). On the existence of n-geodesically complete or future complete solutions of Einstein's field equations with smooth asymptotic structure. *Communications in Mathematical Physics*, 107, 587–609.
- Friedrich, H. (1992). Asymptotic structure of space-time. In A. I. Janis & John R. Porter (Eds.), *Recent advances in general relativity*. Boston: Birkhäuser Inc.
- Friedrich, H. (1995). Einstein equations and conformal structure: Existence of anti-de Sitter-type space-times. *Journal of Geometry and Physics*, 17, 125–184.
- Friedrich, H. (1998a). Einstein's equation and conformal structure. In Huggett et al. (1998).
- Friedrich, H. (1998b). Einstein's equation and geometric asymptotics. In N. Dadhich & J. Narlikar (Eds.), *Gravitation and relativity: At the turn of the millennium. Proceedings of the GR-15 conference*. Pune, India: IUCAA.
- Friedrich, H. (1998c). Gravitational fields near space-like and null infinity. *Journal of Geometry and Physics*, 24, 83–163.
- Friedrich, H. (2002). Conformal Einstein evolution. In J. Frauendiener & H. Friedrich (Eds.), *The conformal structure of space-time: Geometry, analysis, numerics* (Vol. 604, pp. 1–50), Lecture notes in physics. Heidelberg: Springer
- Friedrich, H. (2003). Conformal geodesics on vacuum space-times. *Communications in Mathematical Physics*, 235, 513–543.
- Friedrich, H. (2004). Smoothness at null-infinity and the structure of initial data. In P. Chruściel & H. Friedrich (Eds.), *The Einstein equations and the large scale behaviour of gravitational fields: 50 years of the Cauchy problem in General Relativity*. Basel: Birkhäuser.
- Geroch, R. (1977). Asymptotic structure of space-time. In F. P. Esposito & L. Witten (Eds.), *Asymptotic Structure of space-time*. New York: Plenum Press.
- Huggett, S. A., Mason, L. J., Tod, K. P., Tsou, S. S., & Woodhouse, N. M. J. (Eds.) (1998). *The geometric universe: Science, geometry and the work of roger penrose*. Oxford: Oxford University Press.
- Jordan, P., Ehlers, J., & Sachs, R. K. (1961). Beiträge zur Theorie der reinen Gravitationsstrahlung. *Abhandlungen der Mathematisch-Naturwissenschaftlichen Klasse. Akademie der Wissenschaften und der Literatur in Mainz*, 1, 1–85.
- Kennefick, D. (1997). Controversies in the history of the radiation reaction problem in general relativity. arXiv gr-qc/9704002.
- Klainermann, S., & Nicolò, F. (2003). Peeling properties of asymptotically flat solutions to the Einstein vacuum equations. *Classical and Quantum Gravity*, 20, 3215–3257.
- Lichnerowicz, A. (1958). Sur les ondes et radiations gravitationnelles. *Comptes Rendus de l'Académie des Sciences*, 246, 893–896.
- Lindblad, H., & Rodnianski, I. (2005). Global existence for the Einstein vacuum equations in wave coordinates. *Communications in Mathematical Physics*, 256, 43–110.

- McVittie, G. C. (1955). Gravitational waves and one-dimensional Einsteinian gas-dynamics. *Journal of Rational Mechanics and Analysis*, 4, 201–220.
- Newman, E. T. (1961a). New approach to the Einstein and Maxwell-Einstein field equations. *Journal of Mathematical Physics*, 2, 674–676.
- Newman, E. T. (1961b). Some properties of empty space-times. *Journal of Mathematical Physics*, 2, 324–327.
- Newman, E. T. (1976). Heaven and its properties. *General Relativity and Gravitation*, 7, 107–111.
- Newman, E. T., Kozameh, C. N., & Silva-Ortigoza, G. (2008). On extracting physical content from asymptotically flat spacetime metrics. *Classical and Quantum Gravity*, 25, 145001.
- Newman, E. T., & Penrose, R. (1962). An approach to gravitational radiation by a method of spin coefficients. *Journal of Mathematical Physics*, 3, 566–578. Errata *ibid.* 4 (1963), 998.
- Newman, E. T., & Penrose, R. (1968). New conservation laws for zero rest-mass fields in asymptotically flat space-time. *Proceedings of the Royal Society of London A*, 305, 175–204.
- Newman, E. T., & Unti, T. W. J. (1962). Behavior of asymptotically flat empty spaces. *Journal of Mathematical Physics*, 3, 891–901.
- Penrose, R. (1960). A spinor approach to general relativity. *Annals of Physics*, 10, 171–201.
- Penrose, R. (1963). Asymptotic properties of fields and space-times. *Physical Review Letters*, 10, 66–68.
- Penrose, R. (1964a). Conformal treatment of infinity. In C. DeWitt & B. DeWitt (Eds.), *Relativity, groups and topology*. New York: Gordon and Breach
- Penrose, R. (1964b). The light cone at infinity. In L. Infeld (Ed.), *Relativistic theories of gravitation*. Oxford: Pergamon Press
- Penrose, R. (1965). Zero rest-mass fields including gravitation: Asymptotic behaviour. *Proceedings of the Royal Society of London A*, 284, 159–203.
- Penrose, R. (1999). The central programme of twistor theory. *Chaos, Solitons & Fractals*, 10, 581–611.
- Penrose, R., & Rindler, W. (1986). *Spinors and spacetime* (Vol. 2). Cambridge: Cambridge University Press.
- Pirani, F. A. E. (1957). Invariant formulation of gravitational radiation theory. *Physics Review*, 105, 1089–1099.
- Rosen, N. (1937). Plane polarised waves in the general theory of relativity. *Physikalische Zeitschrift der Sowjetunion*, 12, 366–372.
- Sachs, R. K. (1960). Propagation laws for null and type III gravitational waves. *Zeitschrift für Physik*, 157, 462–477.
- Sachs, R. K. (1961). Gravitational waves in general relativity VI. The outgoing radiation condition. *Proceedings of the Royal Society of London A*, 264, 309–338.
- Sachs, R. K. (1962a). Asymptotic symmetries in gravitational theories. *Physics Review*, 128, 2851–2864.
- Sachs, R. K. (1962b). Gravitational waves in general relativity VIII. Waves in asymptotically flat space-time. *Proceedings of the Royal Society of London A*, 270, 103–127.
- Sachs, R. K. (1964). Characteristic initial value problem for gravitational theory. In L. Infeld (Ed.), *Relativistic theories of gravitation*. Oxford: Pergamon Press.
- Scheidegger, A. E. (1953). Gravitational motion. *Reviews of Modern Physics*, 25, 451–468.
- Schmidt, B. G. (1978). Asymptotic structure of isolated systems. In J. Ehlers (Ed.), *Isolated gravitating systems in general relativity*. New York: Academic Press
- Taub, A. H. (1951). Empty space-times admitting a three parameter group of motions. *Annals of Mathematics*, 53, 472–490.
- Trautman, A. (1958a). Boundary conditions at infinity for physical theories. *Bulletin de l'Académie Polonaise des Sciences III*, 6, 403–406.
- Trautman, A. (1958b). Radiation and boundary conditions in the theory of gravitation. *Bulletin de l'Académie Polonaise des Sciences III*, 6, 407–412.
- Trautman, A. (1958c). On gravitational radiation damping. *Bulletin de l'Académie Polonaise des Sciences III*, 6, 627–633.

Chapter 17

String Theory and Spacetime Geometry



Matthias R. Gaberdiel

17.1 Generalities

In this short article I want to explain some features of string theory, in particular in relation to notions of spacetime geometry. Obviously, given the length of the contribution, I will only be able to give a very cursory overview from a rather idiosyncratic point of view. In particular, there will be many omissions and I will only attempt to sketch one route through some aspects of the theory. I shall also give almost no references and refer for a more comprehensive treatment of string theory to the books by Green et al. (1987) and by Polchinski (1998). A very nice introductory text is also available from Zwiebach (2004).

String theory is, in my opinion, a promising approach towards a theory that combines and incorporates two of the most successful theories of the last century: the standard model of high energy physics that seems to give a very accurate description of the interactions of the elementary particles; and Einstein's general relativity that describes (again very successfully) gravitational effects. The standard model is a quantum theory. It is incomplete as it stands since it ignores the effects of the gravitational force; this is worse than it may sound since the intermediate particles that may be created at sufficiently high energies will cause for example black holes to form. General relativity, on the other hand, is a 'classical' theory. It leads typically (*i.e.* for certain generic classes of initial conditions) to singularities at which the theory ceases to make sense. In order to incorporate the gravitational effects into the standard model (and thereby maybe also cure the problem with

M. R. Gaberdiel (✉)

Institut für Theoretische Physik, ETH Zürich, CH-8093 Zürich, Switzerland

e-mail: gaberdiel@itp.phys.ethz.ch

© Springer Science+Business Media, LLC, part of Springer Nature 2018

D. E. Rowe et al. (eds.), *Beyond Einstein*, Einstein Studies 14,

https://doi.org/10.1007/978-1-4939-7708-6_17

475

the classical singularities of general relativity) it is usually believed that one will have to quantize gravity. Unfortunately, this does not seem to be possible in a straightforward manner since general relativity does not lead to a consistent quantum theory.¹

The starting point of string theory is rather unconventional, and certainly not what you would begin with if you started out with the intention of quantizing gravity. (Indeed, the history of the subject is somewhat convoluted, and I shall not attempt to describe it here.) The basic idea of the theory (as we understand it today) is that the fundamental object in terms of which the theory is being formulated is a one-dimensional string. This string moves initially in a given background geometry (although we shall qualify this statement later). The string has a finite tension—this is essentially the only dimensional parameter in string theory, and it is usually believed to be in the range of the Planck scale—and the classical equations of motion are the relativistic analogue of those equations that govern for example the motions of a violin string. The violin string can vibrate—this is what produces the music—and so does the relativistic string. The different vibrational modes of the string (*i.e.* the analogue of the different notes) are the quanta of the theory, and are thought to correspond to the different ‘elementary particles’ of high energy physics. Strings interact by joining and splitting, and this incorporates the familiar interactions of the elementary particles. String theory is therefore a quantum theory in which the different elementary particles and their interactions arise from one common principle. In addition to the usual elementary particles of high energy physics, the string always possesses one excitation that corresponds to the graviton,² the mode describing the fluctuations of the background geometry in which the string is initially placed. Furthermore, the consistency conditions of the theory require the background geometry to satisfy a generalization of Einstein’s equations of general relativity. One can also calculate the scattering amplitudes of the gravitons, and they are (up to corrections that are suppressed by the string scale) precisely those one would obtain from the Einstein-Hilbert action. Thus string theory defines a consistent quantum theory that combines the quantum field theories of high energy physics with general relativity.

While this sounds very appealing, string theory is certainly not a complete theory, and there are many aspects one still understands rather poorly. In particular, string theory is so far really only a first quantized theory. There is a general belief that a second quantized theory can be formulated by postulating that the ‘Feynman rules’ of the theory arise by taking integrals over surfaces (that ‘fatten out’ the usual Feynman graphs), but there is no fundamental understanding of the underlying

¹There have been recent speculations that $\mathcal{N} = 8$ supergravity in four dimensions may define a (perturbatively) well-defined quantum theory by itself (Bern et al. 2007). There is a long history of such claims and counterclaims, and the situation is certainly not resolved by any means. In any case, $\mathcal{N} = 8$ supergravity is a very special highly symmetric theory that bears little resemblance to the particle spectrum we observe.

²It also contains an infinite tower of massive particles, the lightest of which have Planck scale masses.

principle from which one could derive these generalized Feynman rules. In recent years there have been some impressive successes with this approach to ‘string field theory’—in particular it has been possible to confirm quantitatively (Sen and Zwiebach 2000) Sen’s intuition about D-brane decay—but the general state of affairs is still unsatisfactory. There is also the problem of background independence: the present formulation of the theory seems to require a choice of some background around which one can then study the fluctuations. While there exist convincing arguments that show that the theory is in fact background independent, it would be important to have a formulation in which this is manifest.

Apart from these conceptual problems, there are also important technical problems in applying the theory to situations of interest. In particular, string theory is fairly poorly understood in time-dependent backgrounds. Such backgrounds are obviously crucial for cosmological applications of the theory that probably hold the biggest promise of leading to testable predictions of string theory in the foreseeable future. (Given the smallness of the Planck scale—it is around $10^{-35}m$ —at which effects of quantum gravity are believed to become significant, it will be very difficult to make direct observations of such effects, thus proving or disproving string theory. Incidentally, this problem applies to *any* theory of quantum gravity, not just to string theory.)

On the other hand, there are also spectacular successes that make you (or at least me) believe that the theory has some grain of truth in it. For example, string theory gives one (if not the only) convincing explanation of black hole entropy in terms of microstates (Strominger and Vafa 1996). This comes out right on the nose, without the need to fix any free parameters—indeed there are none!³ There are also deep and detailed results about gauge theories that have come out of the AdS/CFT correspondence, for example predictions for anomalous dimensions of certain operators in $N = 4$ gauge theories to arbitrary order in perturbation theory (Beisert et al. 2007; Bajnok and Janik 2009) that were subsequently confirmed (up to four loops) by gauge theory calculations. (The relevant coefficients are sums of zeta functions with complicated rational coefficients, and it is highly non-trivial that this comes out right!) String theory has given rise to deep insights into various parts of mathematics, in particular, algebraic geometry (mirror symmetry), group theory (monstrous moonshine), number theory (modular forms), to name just a few. At the very least this shows in my opinion very convincingly that the mathematical structure of string theory is very interesting and deep; obviously, none of this implies that the theory must have anything to do with the real world, but even if it should not, it will lead to (and has done so in the past) interesting advances that have been crucial in other contexts.

The way I like to describe the state of affairs is that string theory is a building site: one can discern the rough outline of the building that will be erected in due course; some isolated parts of the building have already been created and decorated in

³In fairness, it should be said that this microstate counting only works well for BPS or near-BPS black holes, although this is something where progress is currently being made.

stunning detail; here and there there are some beautiful sculptures that give witness to the proficiency of the builders, but some connecting parts are still badly missing.

17.2 Bosonic String Theory

In the following I want to sketch some of the basic ideas of string theory in some more detail. To describe the excitations of the string quantitatively it is useful to think of the motion of the string in terms of the two-dimensional surface that is swept out as the one-dimensional string propagates in time through space; in analogy to the world-line of a point particle, this surface is called the *world-sheet* of the string. We can think of the world-sheet as a fixed surface Σ ; the motion of the string is then described by how this fixed surface is embedded in the target space $\mathbb{R}^{1,d}$, *i.e.* by the map $\mathbf{X} : \Sigma \rightarrow \mathbb{R}^{1,d}$. The classical equations of motion of the relativistic string are then determined by the *Nambu-Goto action* (which is simply proportional to the area of the world-sheet in the induced metric)

$$S = -\frac{T}{2} \int d\sigma d\tau \sqrt{(\dot{\mathbf{X}} \cdot \mathbf{X}')^2 - (\dot{\mathbf{X}})^2 (\mathbf{X}')^2}. \quad (17.1)$$

Here $\dot{\mathbf{X}}$ denotes the derivative with respect to the time coordinate τ on the world-sheet, while \mathbf{X}' is the derivative with respect to the world-sheet space coordinate σ , and the inner products are defined with respect to the target space metric (in our case simply the Minkowski metric). Furthermore, T denotes the string tension, which one often parametrizes as $T = \frac{1}{2\pi\alpha'}$. The parameter α' has dimension of length squared—the world-sheet coordinates are here taken to be dimensionless.

The topology of the world-sheet we consider depends on which process we want to analyze—as we mentioned before, at present we only know how to calculate on-shell scattering amplitudes of external string states. For example, if we want to describe the 2-to-2 scattering amplitude, then the world-sheet is the surface that interpolates between two incoming and two out-going string states. There are many surfaces that interpolate between such configurations, and in string field perturbation theory we need to sum (and integrate) over all of them. For the moment we only want to study a single such world-sheet. The simplest situation is that of a single free string; from the point of view of the Feynman rules this describes a single line, and hence corresponds to the ‘propagator’ in field theory.

There are two types of strings, *closed strings* and *open strings*. Closed strings are strings that form closed loops without end-points; from the point of view of the world-sheet the analogue of the propagator then has the topology of a cylinder. One can also consider strings that have two end-points and that are therefore called open strings; for them, the corresponding world-sheet propagator is a rectangle (or a semi-infinite strip). For open strings one also has to specify boundary conditions at the end-points of the open string; this leads to the notion of D-branes that we shall touch upon later on.

The above action is classically equivalent to the theory described by means of the *Polyakov action*

$$S = -\frac{T}{2} \int dx^1 dx^2 \partial_m X^\mu \partial_n X^\nu G_{\mu\nu} \sqrt{-h} h^{mn}, \quad (17.2)$$

where h_{mn} is the metric on the world-sheet (with determinant h)— m and n take the values $m, n = 1, 2$ and refer to world-sheet coordinates—and h^{mn} is the inverse metric. On the other hand G describes the metric in target space. If the target space is just Minkowski space then G is the Minkowski metric; in general, however, the metric G may be a function of the coordinates X . The resulting action is then called a (*non-linear*) *sigma-model*.

Either action is manifestly reparametrization invariant. We can use this reparametrization invariance to go to ‘conformal gauge’ in which we take the world-sheet metric h to be proportional to the usual 2d Minkowski metric. The residual symmetry is then the conformal symmetry, *i.e.* the symmetry that leaves angles invariant (but not necessarily lengths). In this gauge the resulting theory is therefore a conformally invariant (or just conformal) field theory.

In two dimensions conformal symmetry is a very powerful symmetry indeed. If we think of the world-sheet as being a subset of the complex plane,⁴ then *any* analytic function defines a transformation that preserve angles (but not necessarily lengths). We can expand such a function in terms of a Laurent series, and thus the Lie algebra of infinitesimal conformal transformations has a basis of the form

$$L_n \leftrightarrow z^{n+1} \partial_z. \quad (17.3)$$

The Lie algebra of these generators is the so-called Witt algebra. In the quantum theory, a non-trivial central extension of this algebra appears, the famous *Virasoro algebra*

$$[L_m, L_n] = (m - n)L_{m+n} + \frac{c}{12} m(m^2 - 1) \delta_{m, -n}. \quad (17.4)$$

Here c is the central charge; for the situation at hand, c takes the value $c = D = d + 1$, where d is the space-dimension of the target space (taken to be Minkowski space, say).

Because of this very large symmetry, many 2d conformal field theories can be solved exactly, based on symmetry considerations alone. Indeed, the conformal symmetry fixes the structure of the correlation functions up to some constants that can then be determined using factorization and crossing constraints—this is sometimes referred to as the ‘conformal bootstrap’. The representation theory of the Virasoro algebra is interesting, and the algebra of chiral conformal fields define

⁴Here we think of the world-sheet as having Euclidean signature; implicitly we have therefore performed a Wick rotation of the world-sheet theory.

what is called a ‘vertex operator algebra’ in mathematics. This structure has had an important impact in various areas of mathematics, most prominently in the proof of ‘Monstrous Moonshine’ for which Richard Borcherds was awarded the Fields Medal in 1998.

From the point of view of string theory, the conformal symmetry is a gauge symmetry, *i.e.* configurations related by conformal transformations to one another should be thought of as describing the same physical state. In quantizing such a gauge symmetric theory there are two obvious ways in which one may proceed. One may first eliminate the gauge freedom by going to a specific gauge in which the gauge symmetry is completely fixed. Then one determines the classical degrees of freedom and quantizes them. In the context of string theory this procedure is called ‘light-cone gauge’. The other method, called ‘covariant gauge’ in string theory, proceeds by quantizing the gauge invariant theory. Just as in electrodynamics the resulting space of quantum states is not positive (semi-)definite, and one needs to impose the gauge condition after quantization in order to eliminate the states of negative norm. (In the context of electrodynamics this is the Gupta-Bleuler method.)

In string theory, both approaches lead to an interesting constraint on the spacetime dimension. In light-cone gauge one singles out two of the space-time dimensions—one space and one time direction—and identifies them with the space and time coordinate on the world-sheet. This fixes the reparametrization symmetry completely. However, the resulting theory is not manifestly Poincaré invariant (with respect to the Poincaré transformations of the D -dimensional Minkowski space) any longer. After quantization one finds that there is an anomaly in the Poincaré symmetry unless $D = 26$ (Goddard et al. 1973). In covariant gauge one can show that the gauge conditions (the Virasoro conditions) only eliminate the states of negative norm if the central charge of the conformal field theory satisfies (Goddard and Thorn 1972)

$$c \leq 26 . \tag{17.5}$$

The latter condition applies directly also to more general spacetimes rather than just Minkowski space. In fact, one is usually interested in the critical dimension $c = D = 26$ since theories with smaller values of c may be thought of as arising from a theory at $c = D = 26$ upon ‘compactification’. The idea of compactification is that the total 26-dimensional spacetime is a product of some internal compact manifold of dimension d and $(26 - d)$ -dimensional Minkowski space. In particular, for $d = 22$, the resulting macroscopic spacetime is the familiar 4-dimensional Minkowski space.

In general, the internal theory may be defined in terms of a non-linear sigma model, *i.e.* in terms of an action of the form (17.2) but this need not be the case: in order to eliminate the negative norm states in covariant gauge we only need that the corresponding conformal field theory has central charge $c = 22$. In general, it is therefore not really appropriate to talk about an ‘internal manifold’ on which one compactifies the string since the corresponding conformal field theory may not have a geometric interpretation at all. In fact, in the bosonic case we have been discussing

so far, such an interpretation will not be available in general. As we shall see later, the situation is somewhat better in the presence of supersymmetry.

17.3 The Superstring

Up to now we have described the bosonic string. The spacetime excitations it describes are all bosonic; our world clearly also has fermionic degrees of freedom, so we should extend our description to one in which also spacetime fermions can be accommodated. This is best achieved in terms of the so-called superstring in which we not only have spacetime fermions, but also gain an additional symmetry: spacetime supersymmetry that relates fermions and bosons into one another. From the world-sheet point of view this is achieved by introducing also world-sheet fermions in a world-sheet supersymmetric fashion, *i.e.* one for each bosonic direction.⁵ After going to conformal gauge the residual symmetry is then the superconformal algebra that is generated by the generators

$$\begin{aligned} [L_m, L_n] &= (m - n)L_{m+n} + \frac{c}{12}m(m^2 - 1)\delta_{m,-n} \\ [L_m, G_r] &= \left(\frac{m}{2} - r\right)G_{m+r} \\ \{G_r, G_s\} &= 2L_{r+2} + \frac{c}{3}\left(r^2 - \frac{1}{4}\right)\delta_{r,-s} . \end{aligned} \tag{17.6}$$

Again, this string theory can either be quantized in light-cone gauge or covariantly, and the analysis is as before: the condition that the Poincaré symmetry of the light-cone theory is not anomalous requires now $c = 15$. Similarly, the requirement that the negative norm states decouple in the covariant description leads to (compare (17.5))

$$c \leq 15 . \tag{17.7}$$

Since every dimension now contributes $c = 1 + \frac{1}{2}$ to the conformal charge—each boson contributes $c = 1$, while each fermion gives $c = \frac{1}{2}$ —the critical dimension of the superstring (that corresponds to $c = 15$) is $D = 10$.

The fermionic string has world-sheet supersymmetry, but not necessarily spacetime supersymmetry. In fact, there exist a number of non-supersymmetric string theories with spacetime fermions in $D = 10$ dimensions. However, the most interesting (and best understood) string theories are those that possess spacetime

⁵Here we are sketching the RNS construction of the superstring. There exist also other formulations, in particular the manifestly spacetime supersymmetric description due to Green and Schwarz (see for example Green et al. 1987), as well as the more recent pure spinor formulation of Berkovits (Berkovits 2000).

supersymmetry. From the point of view of the world-sheet, spacetime supersymmetry requires that the superconformal field theory has an extended symmetry, namely the $N = 2$ superconformal algebra

$$\begin{aligned}
 [L_m, L_n] &= (m - n)L_{m+n} + \frac{c}{12}m(m^2 - 1)\delta_{m, -n} \\
 [L_m, J_n] &= -nJ_{m+n} \\
 [L_m, G_r^\pm] &= \left(\frac{m}{2} - r\right)G_{m+r} \\
 [J_m, J_n] &= \frac{c}{3}m\delta_{m, -n} \\
 [J_m, G_r^\pm] &= \pm G_{m+r}^\pm \\
 \{G_r^+, G_s^+\} &= \{G_r^-, G_s^-\} = 0 \\
 \{G_r^+, G_s^-\} &= 2L_{r+s} + (r - s)J_{r+s} + \frac{c}{3}\left(r^2 - \frac{1}{4}\right)\delta_{r, -s}.
 \end{aligned} \tag{17.8}$$

Only the $N = 1$ subalgebra that is generated by L_m and $G_r = G_r^+ + G_r^-$ is the residual gauge symmetry that is eliminated by going to light-cone gauge (or for which some gauge fixing condition needs to be imposed in covariant gauge). The additional $N = 2$ generators do not describe gauge symmetries; however, they do constrain the spectrum and the correlation functions of the superconformal field theory, and their presence (together with some quantization condition on the $U(1)$ -charges) is equivalent to the existence of space-time supersymmetry generators.

The analysis so far has been appropriate for open strings (although we have not yet discussed the boundary conditions that need to be imposed—this will be described below). For closed strings there are two independent sets of conformal (or superconformal) symmetry transformations. Indeed the excitations of the closed string can be described in terms of left- and right-moving waves,⁶ and we have symmetries that act separately on them. The symmetries that act on the left-moving excitations will be labelled by L_m , G_r^\pm , and J_n , while those of the right-moving excitations will be denoted by \bar{L}_m , \bar{G}_r^\pm , \bar{J}_n .

17.3.1 Mirror Symmetry

The $N = 2$ superconformal symmetry does not only imply that the theory is actually spacetime supersymmetric. It also allows one to identify the ‘topology of the underlying geometry’ (although there may not be any conventional geometry at all). Indeed, it follows from the above (anti)-commutation relations that either of the two operators

⁶For example, for the case where the world-sheet Σ is a cylinder, these are the waves that travel clockwise or anti-clockwise along the cylinder.

$$Q_{\pm} = G_{-1/2}^{\pm} \quad \text{satisfies} \quad Q_{\pm}^2 = 0. \quad (17.9)$$

The same statement is obviously true for $\bar{Q}_{\pm} = \bar{G}_{-1/2}^{\pm}$. Each of these operators therefore defines a cohomology (Lerche et al. 1989), and we can consider, say, the combined cohomology of Q_+ and \bar{Q}_+ . This is to say, we only consider the subspace of states that are annihilated by Q_+ and \bar{Q}_+ , and we identify states that differ by elements in the image of either Q_+ or \bar{Q}_+ . The resulting space is usually finite-dimensional, and it is called the *topological twist* (Witten 1991). There are obviously four different possibilities, depending on whether one considers Q_+ or Q_- , and \bar{Q}_+ or \bar{Q}_- . However, because Q_+ and Q_- are charge conjugate to one another (and likewise for \bar{Q}_+ and \bar{Q}_-) there are really only two different choices, corresponding, for example, to taking (Q_-, \bar{Q}_+) and (Q_+, \bar{Q}_+) , respectively. The resulting topologically twisted theories are usually called the A-model and B-model, respectively.

For the following it is important that the $N = 2$ superconformal algebra (17.8) has an (outer) automorphism (the *mirror automorphism*) that acts as

$$L_n \mapsto L_n, \quad G_r^{\pm} \mapsto G_r^{\mp}, \quad J_n \mapsto -J_n. \quad (17.10)$$

Since this automorphism exchanges Q_+ and Q_- it induces in particular a one-to-one map between the states of the A-model and that of the B-model. Thus the number of states in the A- and the B-model are the same. However, their spectrum with respect to J_0 and \bar{J}_0 is typically different.

The topological twist has a direct relation to the geometric cohomology of the target space. Suppose that we consider strings propagating on some target manifold M , i.e. a sigma model action of the form (17.2), where the world-sheet fields \mathbf{X} are maps $\mathbf{X} : \Sigma \mapsto M$, and G is the metric on M . Then the geometric Dolbeault cohomology of M , $H_{p,q}(M)$, agrees precisely with the topological twist, where p and q are related to the eigenvalues of J_0 and \bar{J}_0 , respectively. In this sense the topology of the target space manifold can be read off from the string theory description. In particular, given an arbitrary string background we can calculate the topologically twisted theory, and in this manner determine the ‘topology’ of the target space, even if the string theory background was not described in terms of a sigma model in the first place. In this sense we can think of the internal space of any supersymmetric string compactification as having a topology (although it may not have a standard geometrical interpretation).

However, as should be clear from the above discussion, there is now an embarrassment of riches: there is not just one topological twisted theory we can consider, but actually two. Which of the two should we regard as capturing the target space geometry? The answer is surprising and beautiful: both descriptions are equally valid! As we explained above, the A-model and the B-model typically differ by their spectrum of J_0 , \bar{J}_0 eigenvalues, and therefore lead to different Dolbeault cohomologies in general. However, the relation between the two descriptions comes from an automorphism of the $N = 2$ algebra, and hence from an isomorphism of

conformal field theories. Thus the sigma models corresponding to the two different target space geometries define equivalent conformal field theories. Put differently, the string cannot distinguish between the two geometries, and both of them describe equally the target space geometry that is seen by the string.

Geometries that are related to one another in this manner are usually called *mirror manifolds*, and the underlying symmetry is called *mirror symmetry*. Its usefulness in Mathematics stems from the fact that certain calculations are much easier in one description (say the A-model) than the other (say the B-model); by performing a calculation in the A-model one can then make a prediction for certain properties of the mirror partner using mirror symmetry. This was explored by Candelas et al. (1991) for the case of the closed string, and by Walcher (2007) for open strings. The resulting predictions were subsequently proven mathematically using methods of algebraic geometry.

17.3.2 Calabi-Yau Manifolds

The internal manifolds that lead to spacetime supersymmetric string theories in 4 dimensions are the Calabi-Yau manifolds (Candelas et al. 1985) that have the property that they possess a covariantly constant spinor.⁷ The simplest example of a Calabi-Yau manifold is the quintic manifold; at the Fermat point it is described by the equation

$$W_0(x_1, \dots, x_5) = x_1^5 + x_2^5 + x_3^5 + x_4^5 + x_5^5 = 0 \quad (17.11)$$

in complex projective space $\mathbb{C}\mathbb{P}^4$. For this example one also knows the corresponding conformal field theory explicitly (Gepner 1988). There is also a very convenient description of the topological B-model in this case, namely as a Landau-Ginzburg model with superpotential W_0 (Greene et al. 1989).

The above Equation (17.11) only describes a quintic manifold with a special complex structure. In fact, there is a whole moduli space of complex structures within the same topological class; the most general such manifold is described by adding to W_0 an arbitrary homogeneous fifth order polynomial in the variables x_1, \dots, x_5 . Terms that are proportional to x_i^4 can be absorbed into redefining the variables x_i , and hence the non-trivial complex structure deformations are parametrised by the monomials

$$x_i^3 x_j^2, \quad x_i^3 x_j x_k, \quad x_i^2 x_j^2 x_k, \quad x_i^2 x_j x_k x_l, \quad x_1 \cdots x_5 \quad (17.12)$$

of which there are

⁷More recently it has been realized that the condition to preserve supersymmetry in 4 dimensions is slightly weaker, and that also ‘generalized’ Calabi-Yau manifolds (Hitchin 2003; Gualtieri 2004) have this property (Gates et al. 1984).

$$5 \cdot 4 + 5 \cdot \binom{4}{2} + 3 \binom{5}{2} + 5 \cdot 4 + 1 = 20 + 30 + 30 + 20 + 1 = 101. \quad (17.13)$$

Thus the complex structure moduli space is 101-dimensional in this case.

From the point of view of the 4-dimensional spacetime theory, moduli (such as the 101 complex structure moduli of the quintic) describe massless scalars. Such scalars would give rise to long-range forces that are not observed in nature. In order to obtain a phenomenologically viable string description, it is therefore important to find backgrounds without such moduli, *i.e.* to fix (or stabilize) the moduli. It is believed that this can be achieved in the context of generalized Calabi-Yau manifolds that describe backgrounds with fluxes; for a review of the current state of the art see for example Douglas and Kachru (2007) or Blumenhagen et al. (2007).

17.4 D-Branes

Up to now we have mainly considered closed strings for which the string forms a loop without endpoints. As we have mentioned before, we can also have open strings, *i.e.* strings that have two endpoints. Every open string theory contains also closed strings since in interactions the two endpoints of an open string can join to form a closed string. On the other hand, there are closed string theories that do not allow for the inclusion of open strings.

Open strings are characterized by the string action (*i.e.* the Nambu-Goto or the Polyakov action), but we need to specify in addition the relevant boundary conditions. For strings propagating in flat space, the simplest boundary conditions are what are called Neumann or Dirichlet boundary conditions. They require that the endpoint of the open string can either move freely (Neumann), or has a fixed position (Dirichlet). We can fix such a boundary condition separately for the different coordinates of the target space, *i.e.* we can impose Neumann boundary conditions for some directions, and Dirichlet boundary conditions for the others. We can visualize this collection of boundary conditions (for the different directions) by drawing the hypersurface that describes the possible endpoints of the open string. (The tangential directions to the hypersurface are then the directions along which Neumann boundary conditions have been imposed; on the other hand we have Dirichlet boundary conditions for the transversal directions.) This hypersurface is called a D-brane (Polchinski 1995), and the notion of D-branes is also used for backgrounds that are not just flat Minkowski space. In particular, D-branes can wrap certain cycles in the internal manifold, and we can think of them in this manner. However, as we have seen above, the geometry that is seen by the string does not always agree with our classical notion of geometry, and thus this point of view may be misleading. We should therefore study the behavior of D-branes from the actual string point of view, *i.e.* using conformal field theory techniques. This approach has been successfully used in various contexts, see *e.g.* Recknagel and Schomerus

(1998); Brunner et al. (2000); Bergmann and Gaberdiel (1998), but it also has its limitations since explicit conformal field theory descriptions are only available for rather special string backgrounds. It is therefore sometimes more useful to attempt to find a description only for the topologically twisted subsector—this obviously carries less information than the full conformal field theory description, but it is much easier to handle.

So far, we have only explained how to construct the topological subsector in closed string theory, but there is a similar construction for open strings. Because of the boundary conditions of the open string, the left- and right-moving waves are not independent of one another, and one therefore only has one (super)-conformal symmetry. There are essentially two different types of boundary conditions one can impose: one either identifies G^\pm with \tilde{G}^\pm at the boundary, or with \tilde{G}^\mp . The corresponding branes are called B-type or A-type branes, respectively. We can then perform the same cohomological construction as before with respect to the resulting supercharge. In order for this cohomology to fit together with the topological twist of the closed string theory we can only consider A-type branes in the A-model, and B-type branes in the B-model.

17.4.1 Matrix Factorizations

In the context of Calabi-Yau models we mentioned above that the topologically twisted theory can be described in terms of Landau-Ginzburg models whose superpotential W is directly related to the equation characterizing the classical geometry. For such theories there is an elegant characterisation of (B-type) branes in terms of matrix factorizations that is due to Kontsevich (see Kapustin and Li 2003; Brunner et al. 2006; Orlov 2004). Kontsevich proposed that the B-type branes of the Landau Ginzburg theory with superpotential $W(x_i)$ are in one-to-one correspondence with *matrix factorizations* of $W(x_i)$, *i.e.* with matrices $E(x_i)$ and $J(x_i)$ satisfying

$$E(x_i) \cdot J(x_i) = W(x_i) \cdot \mathbf{1}_{r \times r} . \quad (17.14)$$

Here $E(x_i)$ and $J(x_i)$ are $r \times r$ matrices whose entries are polynomials in the variables x_i . The matrices can have arbitrary size (*i.e.* r is arbitrary), but there are a number of identifications. In particular, matrix factorizations that are related to one another by conjugation by a polynomial matrix (whose inverse is also polynomial, *i.e.* has entries that are all polynomials in the x_i) are to be identified. In addition, the matrix factorization $E = W(x_i) \cdot \mathbf{1}$ and $J = \mathbf{1}$ or $E = \mathbf{1}$ and $J = W(x_i) \cdot \mathbf{1}$ are to be regarded as being trivial.

As we explained above, the quintic has a 101-dimensional complex structure moduli space that is described by (17.12). Similarly, there is often a moduli space of D-branes in a fixed background geometry. For example, for the case of the Fermat quintic W_0 in (17.11) we can give a fairly explicit description of one branch of the moduli space (Baumgartl et al. 2007). To this end we define

$$J_1 = x_1 - \eta x_2, \quad J_3 = ax_3 - bx_4, \quad J_5 = ax_5 - cx_4, \quad (17.15)$$

where η is a fifth root of -1 , and a, b and c are complex numbers. We then look for common solutions of

$$J_1 = J_3 = J_5 = 0, \quad \text{and} \quad W_0 = 0. \quad (17.16)$$

For $J_1 = 0$ we have $x_1 = \eta x_2$, and since η is a fifth root of -1 , this leads to $x_1^5 + x_2^5 = 0$. Let us assume for the moment that $a \neq 0$. Then we can solve $J_3 = 0$ by setting $x_3 = \frac{c}{a}x_4$, and similarly $J_5 = 0$ by $x_5 = \frac{c}{a}x_4$. This then solves $W_0 = 0$ provided that

$$a^5 + b^5 + c^5 = 0. \quad (17.17)$$

Thus we have a common solution to (17.16) provided that (17.17) holds. The Nullstellensatz then implies that we can write

$$W_0 = J_1 \cdot E_1 + J_3 \cdot E_3 + J_5 \cdot E_5, \quad (17.18)$$

where E_1, E_3 and E_5 are all homogeneous polynomials of degree 4 in x_i . By placing these polynomials in appropriate matrix entries we can then construct a matrix factorization of W_0 (Baumgartl et al. 2007). This matrix factorization is also available for $a = 0$, provided that either $b \neq 0$ or $c \neq 0$. Thus the above reasoning shows that the moduli space of B-type D-branes contains a branch that is parametrized by the complex curve (17.17) in complex projective space $\mathbb{C}\mathbb{P}^2$. Each point in this moduli space corresponds to a D-brane; we can think of the D-brane corresponding to (a, b, c) as wrapping the 2-cycle (inside the quintic) described by

$$(x_1, x_2, x_3, x_4, x_5) = (u, \eta u, av, bv, cv), \quad (17.19)$$

where $(u, v) \in \mathbb{C}\mathbb{P}^1$ (Ashok et al. 2004).

As for the closed string moduli discussed before, the open string moduli also describe massless bosons in 4 dimensions that are phenomenologically unacceptable. One is therefore interested in configurations where all, closed and open moduli, are lifted (fixed). Often the two problems—namely fixing the closed and the open moduli—are treated independently from one another, but obviously these moduli spaces are interrelated. In particular, the structure of the moduli space of D-branes depends on precisely which closed string background one considers.⁸ For example, if we change the closed string background by adding to W_0

⁸The D-branes also have a backreaction on the closed string background in which they are placed. This effect, however, only appears at higher order in string perturbation theory (Fischer and Susskind 1986a,b; Keller 2007).

$$W = W_0 + \lambda \Phi, \quad \Phi = x_1^3 \cdot (t_3 x_3^2 + t_4 x_4^2 + t_5 x_5^2), \quad (17.20)$$

then the only matrix factorizations within a neighborhood of this family that are also solutions of the deformed matrix factorization condition are those that satisfy (Baumgartl et al. 2007)

$$a^5 + b^5 + c^5 = 0 = t_3 a^2 + t_4 b^2 + t_5 c^2. \quad (17.21)$$

By Bezout's theorem, there are then only ten discrete solutions. This actually ties in with expectations from geometry: at a generic point in the complex structure moduli space there are only finitely many holomorphic 2-cycles (around which supersymmetric D-branes can wrap). A similar analysis can also be performed for other Calabi-Yau backgrounds (Baumgartl and Wood 2008).

One can also understand dynamically what happens when the closed string background is modified. This question can be analyzed using techniques of conformal perturbation theory (see in particular Fredenhagen et al. 2007). The perturbation that corresponds to the complex structure deformation (17.20) generically breaks the $N = 1$ superconformal symmetry in the presence of a D-brane. This induces a renormalisation group flow that drives the D-brane to one that is compatible with the deformed closed string background. Incidentally, the RG flow is the gradient flow of the effective spacetime superpotential that one can thus calculate in detail (Baumgartl et al. 2007; Baumgartl and Wood 2008).⁹ The effective spacetime superpotential describes the interactions of the massless states in the resulting 4 dimensional spacetime theory, and thus characterizes how open and closed moduli interact. It is therefore of direct phenomenological significance.

17.5 Conclusions

String theory is a promising candidate for a theory unifying the standard model of high energy physics and Einstein's general relativity. According to our current understanding the theory is initially formulated in a fixed background, but it contains the quanta that describe the fluctuations of this background. Among other things, string theory therefore defines a sensible quantum theory of gravity.

The background in which the string propagates does not necessarily have a direct interpretation in terms of classical geometry. In particular, the string background is really encoded in a conformal field theory of an appropriate central charge which may not be equivalent to a sigma model. Even if it *does have* a geometric interpretation, this will typically not be unique since there is at least the ambiguity

⁹Subsequently, this effective spacetime superpotential was reproduced by Aganagic and Beem (2011) using string duality arguments, see also Baumgartl et al. (2011). The same method was also applied to another interesting example by Baumgartl et al. (2012).

associated to mirror symmetry: superstring theory on manifolds that are mirror partners of each other are equivalent to one another. The geometry that is seen by the string—this is sometimes called ‘quantum geometry’—is therefore different from our classical notions of geometry. Gaining insights into the nature of quantum geometry continues to be an important and interesting problem. D-branes (that encode the boundary conditions of open string sectors) provide an interesting new tool for this.

Acknowledgements I thank my collaborators Marco Baumgartl, Ilka Brunner, Stefan Fredenhagen and Christoph Keller for enjoyable collaborations on which some of this work is based. My research has been partially supported by the Swiss National Science Foundation, as well as the Marie Curie network ‘Constituents, Fundamental Forces and Symmetries of the Universe’ (MRTN-CT-2004-005104).

References

- Aganagic, M., & Beem, C. (2011). The geometry of D-brane superpotentials. *Journal of High Energy Physics*, 1112, 060. arXiv:0909.2245 [hep-th].
- Ashok, S. K., Dell’Aquila, E., Diaconescu, D. E., & Florea, B. (2004). Obstructed D-branes in Landau-Ginzburg orbifolds. *Advances in Theoretical and Mathematical Physics*, 8, 427. arXiv:hep-th/0404167.
- Bajnok, Z., & Janik, R. A. (2009). Four-loop perturbative Konishi from strings and finite size effects for multiparticle states. *Nuclear Physics B*, 807, 625. arXiv:0807.0399 [hep-th].
- Baumgartl, M., Brunner, I., & Gaberdiel, M. R. (2007). D-brane superpotentials and RG flows on the quintic. *Journal of High Energy Physics*, 0707, 061. arXiv:0704.2666 [hep-th].
- Baumgartl, M., Brunner, I., & Plencner, D. (2012). D-brane moduli spaces and superpotentials in a two-parameter model. *Journal of High Energy Physics*, 1203, 039. arXiv:1201.4103 [hep-th].
- Baumgartl, M., Brunner, I., & Soroush, M.: D-brane superpotentials (2011). Geometric and worldsheet approaches. *Nuclear Physics B*, 843, 602. arXiv:1007.2447 [hep-th].
- Baumgartl, M., & Wood, S. (2008). Moduli webs and superpotentials for five-branes. *Journal of High Energy Physics*, 0906, 052. arXiv:0812.3397 [hep-th].
- Beisert, N., Eden, B., & Staudacher, M. (2007). Transcendentality and crossing. *Journal of Statistical Mechanics: Theory and Experiment*, 0701, P021. arXiv:hep-th/0610251.
- Bergman, O., & Gaberdiel, M. R. (1998). Stable non-BPS D-particles. *Physics Letters B*, 441, 133. arXiv:hep-th/98055.
- Berkovits, N. (2000). Super-Poincare covariant quantization of the superstring. *Journal of High Energy Physics*, 0004, 018. arXiv:hep-th/0001035.
- Bern, Z., Carrasco, J. J., Dixon, L. J., Johansson, H., Kosower, D. A., & Roiban, R. (2007). Three-Loop superfiniteness of N=8 supergravity. *Physical Review Letters*, 98, 161303. arXiv:hep-th/0702112.
- Blumenhagen, R., Kors, B., Lüst, D., & Stieberger, S. (2007). Four-dimensional string compactifications with D-branes, orientifolds and fluxes. *Physics Reports*, 445, 1. arXiv:hep-th/0610327.
- Brunner, I., Douglas, M. R., Lawrence, A. E., & Römelsberger, C. (2000). D-branes on the quintic. *Journal of High Energy Physics*, 0008, 015. arXiv:hep-th/9906200.
- Brunner, I., Herbst, M., Lerche, W., & Scheuner, B. (2006). Landau-Ginzburg realization of open string TFT. *Journal of High Energy Physics*, 0611, 043. arXiv:hep-th/0305133.
- Candelas, P., De La Ossa, X. C., Green, P. S., & Parkes, L. (1991). A pair of Calabi-Yau manifolds as an exactly soluble superconformal theory. *Nuclear Physics B*, 359, 21.

- Candelas, P., Horowitz, G. T., Strominger, A., & Witten, E. (1985). Vacuum configurations for superstrings. *Nuclear Physics B*, 258, 46.
- Douglas, M. R., & Kachru, S. (2007). Flux compactification. *Reviews of Modern Physics*, 79, 733. arXiv:hep-th/0610102.
- Fischler, W., & Susskind, L. (1986a). Dilaton tadpoles, string condensates and scale invariance. *Physics Letters B*, 171, 383.
- Fischler, W., & Susskind, L. (1986b). Dilaton tadpoles, string condensates and scale invariance 2. *Physics Letters B*, 173, 262.
- Fredenhagen, S., Gaberdiel, M. R., & Keller, C. A. (2007). Bulk induced boundary perturbations. *Journal of Physics A*, 40, F17. arXiv:hep-th/0609034.
- Gates, S. J., Hull, C. M., & Rocek, M. (1984). Twisted multiplets and new supersymmetric nonlinear sigma models. *Nuclear Physics B*, 248, 157.
- Gepner, D. (1988). Space-time supersymmetry in compactified string theory and superconformal models. *Nuclear Physics B*, 296, 757.
- Goddard, P., Goldstone, J., Rebbi, C., & Thorn, C. B. (1973). Quantum dynamics of a massless relativistic string. *Nuclear Physics B*, 56, 109.
- Goddard, P., & Thorn, C. B. (1972). Compatibility of the dual pomeron with unitarity and the absence of ghosts in the dual resonance model. *Physics Letters B*, 40, 235.
- Green, M. B., Schwarz, J. H., & Witten, E. (1987). *Superstring theory I & II*. Cambridge: Cambridge University Press.
- Greene, B. R., Vafa, C., & Warner, N. P. (1989). Calabi-Yau manifolds and renormalization group flows. *Nuclear Physics B*, 324, 371.
- Gualtieri, M. (2004). *Generalized complex geometry*. Oxford University DPhil thesis. arXiv:math.DG/0401221.
- Hitchin, N. (2003). Generalized Calabi-Yau manifolds. *The Quarterly Journal of Mathematics*, 54, 281–308. arXiv:math.DG/0209099.
- Kapustin, A., & Li, Y. (2003). D-branes in Landau-Ginzburg models and algebraic geometry. *Journal of High Energy Physics*, 0312, 005. arXiv:hep-th/0210296
- Keller, C. A. (2007). Brane backreactions and the Fischler-Susskind mechanism in conformal field theory. *Journal of High Energy Physics*, 0712, 046. arXiv:0709.1076 [hep-th].
- Lerche, W., Vafa, C., & Warner, N.P. (1989). Chiral rings in $N=2$ superconformal theories. *Nuclear Physics B*, 324, 427.
- Orlov, D. (2004). Triangulated categories of singularities and D-branes in Landau-Ginzburg models. *Proceedings of the Steklov Institute of Mathematics*, 3(246), 227–248. arXiv:math/0302304.
- Polchinski, J. (1995). Dirichlet-branes and Ramond-Ramond charges. *Physical Review Letters*, 75, 4724. arXiv:hep-th/9510017.
- Polchinski, J. (1998). *String theory I & II*. Cambridge: Cambridge University Press.
- Recknagel, A., & Schomerus, V. (1998). D-branes in Gepner models. *Nuclear Physics B*, 531, 185. arXiv:hep-th/9712186.
- Sen, A., & Zwiebach, B. (2000). Tachyon condensation in string field theory. *Journal of High Energy Physics*, 0003, 002. arXiv:hep-th/9912249.
- Strominger, A., & Vafa, C. (1996). Microscopic origin of the Bekenstein-Hawking entropy. *Physics Letters B*, 379, 99. arXiv:hep-th/9601029.
- Walcher, J. (2007). Opening mirror symmetry on the quintic. *Communications in Mathematical Physics*, 276, 671. arXiv:hep-th/0605162.
- Witten, E. (1991). Mirror manifolds and topological field theory. In S. T. Yau (Ed.), *Mirror symmetry I* (p. 121). Providence, RI: American Mathematical Society. arXiv:hep-th/9112056.
- Zwiebach, B. (2004). *A first course in string theory*. Cambridge: Cambridge University Press.

Conference Program

Volker Bach and David E. Rowe

Conference Program

Beyond Einstein: Historical Perspectives on Geometry, Gravitation, and Cosmology in the Twentieth Century

Johannes Gutenberg-Universität Mainz, 22–26 September, 2008

Session 1: General Relativity, Cosmology, and Unification

Frank Steiner (Ulm): Do we live in a small Universe?

Jim Ritter (Paris): Mathematicians, Einstein, and the Unification Project: A Tale of Two Cities

Hubert Goenner (Göttingen): Unified Field Theory up to the 1960s: its development and the interaction among research groups

Session 2: Formative Ideas for Spacetime Structures

Scott Walter (Nancy): How did Minkowski discover spacetime?

Engelbert Schücking (New York): What is General Relativity?

Harvey Brown (Oxford): Why do rods contract? Extending John Bell's 'Lorentzian pedagogy' into general relativity

Session 3: Relativity in Göttingen and Beyond

Tilman Sauer (Pasadena, California): General relativity from Hilbert's perspective

Katherine Brading (Notre Dame): Hilbert and Einstein's General Theory of Relativity: Two Communications on the Foundations of Physics

Jean Eisenstaedt (Paris): From the Schwarzschild singularity to the black-hole horizon

Session 4: Testing General Relativity and Rival Theories

Clifford Will (St. Louis): Putting General Relativity to the Test: 20th Century Highlights and 21st Century Prospects

V. Bach · D. E. Rowe (✉)
Mainz University, Mainz, Germany
e-mail: rowe@mathematik.uni-mainz.de

Herbert Pfister (Tübingen): Rotating hollow and full spheres: Einstein, Thirring, Lense, and beyond

Allan Franklin (Boulder, Colorado): The rise and fall of the Fifth Force

Session 5: Cycles, Waves, and Inflation

Helge Kragh (Aarhus): Continual Fascination: Oscillatory Cosmological Models after Einstein

Dan Kennefick (Fayetteville, Arkansas): Relativistic Lighthouses: The role of binary pulsars in proving the existence of gravitational waves

Christopher Smeenk (London, Ontario): Inflation as a Theory of Structure Formation

Session 6: Conformal Boundaries and Quantum Geometry

Abhay Ashtekar (Penn State): Classical Singularities and Quantum Spacetime

Jörg Frauendiener (Tübingen): Development and applications of conformal infinity

Roger Penrose (Oxford): Conformal boundaries, Quantum Geometry, and Cyclic Cosmology

Session 7: Mainstream and Exotic Cosmology

Norbert Straumann (Zurich): Problems with modified theories of gravity, as alternatives to dark energy

Hans Jörg Fahr (Bonn): Cosmologies with cosmic vacuum energy decay and the creation of effective cosmic matter

Erhard Scholz (Wuppertal): Can scale covariant Weylian geometry be relevant in contemporary cosmology?

Session 8: Mathematical Motifs in General Relativity

Sergiu Klainerman (Princeton): On Cosmic Censorship and the Cauchy Problem

Donal O'Shea (Mt. Holyoke, Mass.): The Unexpected Resolution of the Poincaré Conjecture

Domenico Giulini (MPI für Gravitationsphysik, Golm): On the Nature of Geometrodynamics

Session 9: Quantum Gravity and String Theory

Robert Wald (Chicago): On Quantum Field Theory in Curved Spacetimes

Matthias Gaberdiel (Zurich): String Theory and Spacetime Geometry

Jürg Fröhlich (Zurich): Events in Quantum Theory and the Emergence of Space-Time