BANGS, CRUNCHES, WHIMPERS, AND SHRIEKS

Singularities and Acausalities in Relativistic Spacetimes

JOHN EARMAN

New York Oxford
OXFORD UNIVERSITY PRESS
1995

Oxford University Press

Oxford New York
Athens Auckland Bangkok Bombay
Calcutta Cape Town Dar es Salaam Delhi
Florence Hong Kong Istanbul Karachi
Kuala Lumpur Madras Madrid Melbourne
Mexico City Nairobi Paris Singapore
Taipei Tokyo Toronto

and associated companies in
Berlin Ibadan

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016

Oxford is a registered trademark of Oxford University Press

Library of Congress Cataloging-in-Publication Data
Earman, John.
Bangs, crunches, whimpers, and shrieks : singularities and acausalities in relativistic spacetimes / John Earman.
p. cm.
Includes bibliographical references and index.
ISBN 0-19-509591-X
1. Space and time. 2. General relativity (Physics)
3. Singularities (Mathematics) I. Title.
QC173.59.S65E15 1995
530.1'1—dc20 95-49132

Printing: 1 3 5 7 9 8 6 4 2
Printed in the United States of America
on acid-free paper

For David, Bob², John², Larry, Tim,
Roberto and all the others
who have sustained me in my labors

PREFACE

The editor of *Philosophy of Science*, the official journal of the Philosophy of Science Association, informs me that in recent years only a small fraction of the submitted papers deal with the foundations of general relativity theory, and of these only a few focus on issues connected with spacetime singularities. This is not because the journal discourages technical papers or papers concerned with foundational issues in physics; indeed, it routinely publishes highly technical articles on the problems of measurement, hidden variables, and non-locality in quantum mechanics. Part of the difference may lie in the fact that while anyone who can understand the notion of a vector space can be brought to appreciate the measurement problem in quantum mechanics and while even undergraduates can derive versions of Bell's inequalities, the analysis of spacetime singularities involves comparatively difficult mathematics. But this cannot be the whole answer since philosophers of science are not strangers to differential geometry. Rather I think that the relative neglect of this area lies in the failure of philosophers to appreciate the seriousness of the foundational issues posed by singularities in general relativity and the importance of these issues for the philosophy of space and time and for the philosophy of science in general.

The present book is dedicated to the goal of ending the neglect. It is aimed primarily at philosophers who have some prior acquaintance with relativity theory and secondarily at philosophically minded physicists. It makes no pretense at being a comprehensive survey of the subject. I doubt that any philosopher at work today is capable of coping with all of the history, philosophy, physics, and mathematics needed to produce such a survey; and even if there were such a polymath, the resulting work would likely be an indigestible tome that would gather dust on the shelves of the few libraries that could afford to buy it. Rather than strive for comprehensiveness, my strategy is to cover some core topics and sufficiently many other topics so that the reader will not be overwhelmed but can nevertheless get a good feel for the topography. The level of presentation is not mathematically rigorous, but I hope it is rigorous enough to illuminate the relevant technical issues in general relativity and to reveal their connection with philosophical issues about the nature of space, time, causality, and the laws of nature. The chapters are interlinked but are designed to be self-contained enough that the reader who tires of one topic, without too much effort, will be able to turn to sampling another. Whether or not readers agree with my analyses and positions, I will be content if they put down this book believing that if

philosophers were doing their jobs in trying to fathom the implications of spacetime singularities, the result would be one of the more exciting areas in the philosophy of science.

For those readers who want to consult background references in the relavant branches of mathematics and physics, I have several suggestions. There are many good textbooks on general relativity. For present purposes, Robert Wald's *General Relativity* (1984a) is especially recommended. Hawking and Ellis's classic, *The Large Scale Structure of Space-Time* (1973), though now out of date, is an indispensable source of key results on singularities. A semipopular but very insightful treatment of several of the topics studied here is to be found in Kip Thorne's *Black Holes and Time Warps* (1994). The most recent technical treatises on singularities are P. S. Joshi's *Global Aspects in Gravitation and Cosmology* (1993) and C. J. S. Clarke's *Analysis of Spacetime Singularities* (1993).

Chapter 3 was based on two of my articles: "The Cosmic Censorship Hypothesis," in J. Earman, A. I. Janis, G. J. Massey, and N. Rescher (eds.), *Philosophical Problems of the Internal and External Worlds: Essays on the Philosophy of Adolf Grünbaum* (Pittsburgh: University of Pittsburgh Press, 1993), and "Cosmic Censorship," *PSA 1992*, Vol. II (East Lansing: Philosophy of Science Association, 1993). Most of the material in Chapter 4 was coauthored by John Norton and appeared as "Forever Is a Day: Super-tasks in Pitowsky and Malament–Hogarth Spacetimes," *Philosophy of Science*, 60 (1993) 22–42. Chapter 6 was based on "Recent Work on Time Travel," in Steven Savitt (ed.), *Time's Arrow Today* (Cambridge: Cambridge University Press, 1994). I am grateful to the editors of these books and to the presses and journals for their kind permission to reprint this material.

Parts of this book were written while I was a participant in the research group on semantical aspects of spacetime theories at the Zentrum für Interdisciplinare Forschung (ZiF), Universität Bielefeld. I am grateful to Ulrich Majer and Heinz-Jürgen Schmidt for organizing the group and to the ZiF for its financial support.

I am indebted to a number of people for ideas and helpful suggestions. At the risk of failing to mention some of them, I would like to thank Chris Clarke, Jean Eisenstaedt, John Norton, and John Stachel. Special thanks are due to David Malament, Tim Maudlin, Roberto Torretti, and Robert Wald for detailed comments on an earlier draft. Andrew Backe provided invaluable help in preparing the bibliography

J.E.

Pittsburgh, Pennsylvania
August 1994

CONTENTS

BANGS, CRUNCHES, WHIMPERS, AND SHRIEKS

1

Introducing Spacetime Singularities and Acausalities

1.1 Introduction

This book is simultaneously an essay in the philosophy of science and the foundations of physics. It seeks to assess the implications of spacetime singularities and acausal features of relativistic spacetimes for the general theory of relativity and the philosophy of space and time. To the extent that it succeeds, it should succeed in convincing the reader that for the constellation of issues to be studied, physics, foundations of physics, and the philosophy of science are all of a piece. Before trying to describe the issues, it may be helpful to begin by setting them in context.

The two most fundamental theories of twentieth century physics, quantum mechanics (QM) and the general theory of relativity (GTR), stand in wary regard of one another. Each is spectacularly successful in its own domain of application, both in making accurate predictions and in providing conceptual understanding of otherwise puzzling phenomena. In the latter achievements lie the roots of what is arguably the most important challenge of contemporary physics. It is hard to see how any future theory that incorporates the successes of both QM and GTR can dispense with the conceptual apparatus of either. But combining them into a unified quantum theory of gravity has proved to be formidable and is thus far unfulfilled.

Despite their successes, QM and GTR are beset by problems that raise worries about the foundations of these theories—QM by the measurement problem and associated problems of non-locality, GTR by the problems that are the focus of this study. Some physicists harbor the hope that both sets of problems will be resolved by the sought-after quantum theory of gravity. It is difficult to assess this hope since we can now only dimly perceive the shape that a successful marriage of QM and GTR will take. And even if the hope is eventually realized, it is important to pursue these foundational problems for the light they may shed on the correct path towards the marriage.

For many purposes, the measurement problem in QM can be ignored by

experimentalists and theoreticians alike. Not surprisingly, it was ignored by a large segment of the physics community, and the opinion was once widespread that this problem is merely a *Scheinproblem*. I vividly recall the occasion of a lecture on the measurement problem given in the early 1970s at The Rockefeller University by a Nobel laureate in physics. The reaction of the audience, composed largely of theoretical physicists and mathematicians, was distinctly cool if not unfriendly. The skepticism was directed not so much at the proposed solution as to the notion that there was a problem to be solved. After the lecture, the laureate remarked ruefully: "I suppose that I will have to do something new to restore my reputation." Today his lecture would likely get a different reception, at least judging from the fact that *The Physical Review*, the most prestigious journal in theoretical physics, now routinely publishes articles on this topic. The implied change in attitude reflects a recognition that the measurement problem poses a fundamental challenge for QM, although how to state the challenge is controversial. On the one hand, the problem can be seen as revealing that there is something rotten at the core of the theory because of its inability to give a satisfactory description of what occurs in the interaction between an object system and a measurement apparatus. Those who share this diagnosis see a need for a new dynamics for QM, involving a non-linear, non-unitary, stochastic evolution to replace the linear, unitary, and deterministic evolution implied by the Schrödinger equation. On the other hand, the challenge can be seen not as calling forth new physics but rather a new understanding of quantum ontology and the ways in which the Hilbert space apparatus of the theory relates to that ontology. Reactions to this way of reading the challenge range from new interpretation rules for assigning values to quantum observables, even when the state vector of the system is not an eigenstate of the operator corresponding to the observable, to a world picture not unlike Borges' garden of forking paths (many worlds interpretation), to a model of reality that replaces many worlds with many minds.[1]

My own conviction is that neither clever semantical rules nor extravagant metaphysics will suffice and that eventually new physical principles will have to be recognized. But this is not the place to argue for that opinion.

The history of the problem of spacetime singularities in GTR has followed a rather different trajectory. As will be seen in the next section, the recognition of the problem came not long after Einstein completed the theory in 1915–16. Over the next four decades the discussion of the problem was somewhat desultory, but no more so than the discussion of relativistic gravitation in general, which came to be regarded as a backwater of physics research. The renaissance of GTR in the 1960s made singularities a focus of attention. The new mathematical techniques developed to analyze solutions of the Einstein gravitational field equations permitted the proof of theorems which showed unequivocally that spacetime singularities could not be ignored but are generic features of general relativistic spacetimes.[2] But this is getting ahead of the story.

1.2 Spacetime singularities: In the beginning

In this section and the following section I will provide a brief sketch of the early struggles with the problem of spacetime singularities, concentrating mainly on Einstein's evolving attitude towards the problem. The selection of topics is guided by the unabashedly Whiggish aim of motivating later chapters, where modern attempts to understand the nature and implications of spacetime singularities in GTR are discussed. A careful history of the subject will have to await another occasion,[3] but nevertheless, I hope that the present account will convey some sense of just how difficult and at times baffling the subject was for the pioneers of GTR. This last remark is not a piece of Whiggish condescension since, as the subsequent chapters will demonstrate, the subject continues to challenge and baffle.

The problem of spacetime singularities in GTR arose even before the theory was put in its final form. As of 18 November 1915, Einstein had not arrived at the final form of his gravitational field equations, although the equations he presented in his communication of that date (Einstein 1915) to the Berlin Academy did reduce to his final equations, hereafter referred to as the Einstein field equations (EFE), in the case of the exterior field of a mass distribution.[4] Einstein produced an approximate solution to these equations for a static spherically symmetric field. Rewritten in spherical coordinates, the line element of this approximate solution is

$$ds^2 = (1 + \alpha/r)\,dr^2 + r^2(d\theta^2 + \sin^2\theta\,d\phi^2) - (1 - \alpha/r)c^2\,dt^2 \qquad (1.1)$$

where $\alpha = 2MG/c^2$, M being the mass of the source, G the gravitational constant, and c the velocity of light.[5] (From here on I will work in units where $c = 1$. A frequent choice of units sets $\alpha = 2M$. Thus, the "Schwarzschild radius" is often designated as $r = 2M$.) One naturally wonders about the meaning of the non-regular points $r = 0$ and $r = \alpha$ where the line element (1.1) ceases to make sense. In November of 1915 Einstein did not have time to ponder such questions, for he was fully occupied in using (1.1) to resolve the long-standing anomaly of the perihelion motion of Mercury.[6] The following year Karl Schwarzschild (1916) produced the exact solution of which (1.1) can be regarded as a first-order approximation. In the coordinates later introduced by Droste (1917), the Schwarzschild line element is

$$ds^2 = (1 - \alpha/r)^{-1}\,dr^2 + r^2(d\theta^2 + \sin^2\theta\,d\phi^2) - (1 - \alpha/r)\,dt^2 \qquad (1.2)$$

Because (1.2) is an exact solution, the question of singularities became more pointed, and was to become still more pointed when Birkhoff (1923) showed that the Schwarzschild solution is the only spherically symmetric solution to the vacuum EFE (with vanishing cosmological constant).

The question was broached by David Hilbert (1917) in Part II of his seminal article "Grundlagen der Physik." Hilbert supplied a general definition

of what it means for the spacetime metric to be non-singular or, to use Hilbert's own terminology, "regular," and he applied his definition to (1.2).

> For $\alpha \neq 0$, it turns out that $r = 0$ and (for positive α) also $r = \alpha$ are points at which the line element is not regular. By that I mean that a line element or a gravitational field $g_{\mu\nu}$ is regular at a point if it is possible to introduce by a reversible, one–one transformation a coordinate system, such that in this coordinate system the corresponding functions $g'_{\mu\nu}$ are regular at that point, i.e., they are continuous and arbitrarily differentiable at the point and in a neighborhood of the point, and the determinant g' is different from 0. (Hilbert 1917, p. 70–71)[7]

Given Hilbert's definition, his conclusion that the Schwarzschild metric is singular at both $r = 0$ and $r = \alpha$ is perfectly correct. But by modern lights the definition is defective in failing to capture the distinction between genuine singularities—as in the case of $r = 0$—and mere coordinate singularities—as in the case of $r = \alpha$. The Droste coordinates in (1.2) give the false appearance of a singularity at $r = \alpha$ because of the peculiar way in which they fail to cover the spacetime manifold. The failure can be made transparent by suppressing the angular coordinates θ and ϕ and by transforming from the r, t coordinates to new coordinates X, T in which the metric components are well behaved everywhere $r > 0$:

$$(r/\alpha - 1)\exp(r/\alpha) = X^2 - T^2$$
$$\frac{t}{\alpha} = \ln\left(\frac{T+X}{X-T}\right) \tag{1.3}$$

The r–t part of the Droste–Schwarzschild line element is replaced by

$$ds^2 = 4\alpha^3 \exp(-r/\alpha)\,(dX^2 - dT^2) \tag{1.4}$$

which is evidently well behaved at $r = \alpha$. The Schwarzschild radius $r = \alpha$ corresponds to the null lines $X = \pm T$. At these locations $t = \pm\infty$, indicating the failure of the t coordinate. Hilbert's demand that regular metric components $g'_{\mu\nu}$ be produced from $g_{\mu\nu}$ by means of a transformation that is smooth and invertible at $r = \alpha$ is thus inappropriate.[8] This little piece of hindsight wisdom did not find its final fruition until forty-three years later when Kruskal (1960) and Szekeres (1960) constructed the maximal analytic extension of the Schwarzschild metric.

Einstein's initial worry about the Schwarzschild solution was not so much with its apparently singular nature as with its anti-Machian character. Rightly or wrongly, Einstein interpreted Mach's principle as implying that the metrical structure of spacetime should be determined by its matter content, and prima facie the Schwarzschild solution is incompatible with this reading of the principle. As John Stachel has written, it was for Einstein a

"scandal that a solution to his field equations should exist which corresponded to the presence of a single body in an otherwise 'empty' universe" (Stachel 1979, p. 440). The scandal erupted anew the following year with the de Sitter solution to EFE. In attempting to quash this second form of the scandal, Einstein was forced to confront the issue of spacetime singularities.

In the coordinates de Sitter (1917a,b) used, the line element of his spacetime takes the form

$$ds^2 = dr^2 - R^2 \sin^2\left(\frac{r}{R}\right)(d\psi^2 + \sin^2\psi\, d\theta^2) - \cos^2\left(\frac{r}{R}\right)dt^2 \tag{1.5}$$

where R is a positive constant. The metric in (1.5) can beconsidered to be a solution to the vacuum EFE with positive cosmological constant $\Lambda = 3/R^2$. In a postcard dated 24 March 1917, Einstein complained to de Sitter that "your solution corresponds to no physical reality."[9] A lengthy correspondence ensued,[10] and after several changes of position Einstein published his complaint in the *Proceedings of the Berlin Academy* (Einstein 1918). The published objection was based on the idea that the de Sitter metric is singular. We can see in Einstein's objection some of the key elements that figure in the modern definition of that term. For the Einstein of 1918, a spacetime is non-singular if "in the finite realm" the covariant and contravariant components of the metric are continuous and differentiable and, consequently, the determinate $g =: \det(g_{ab})$ never vanishes. A point p is said to lie in the finite realm just in case it can be joined to an arbitrarily chosen origin point p_0 of spacetime by a curve whose proper length $\int_{p_0}^{p} ds$ is finite.[11] For the de Sitter solution (1.5), g vanishes at $r = 0$ and $\psi = 0$. According to Einstein, however, this represents only an apparent violation of continuity since $g \neq 0$ can be restored at these points by a change of coordinates. But the discontinuity at $r = \pi R/2$ was another matter. "It seems," Einstein wrote, "that this discontinuity cannot be removed by any choice of coordinates" (Einstein 1918, p. 271). Thus, it seemed to Einstein that "until proof to the contrary" one should conclude that a singularity occurs at these locations since they lie at finite distances—starting at $r = t = 0$ and tracing to $r = \pi R/2$ along a curve with constant ψ, θ, t, the integral $\int_0^{\pi R/2} ds$ is finite.

Although Einstein's ideas here are true to the spirit of the definitions of spacetime singularities that would be developed half a century later, the attempted implementation of the ideas is seen in the harsh glare of hindsight wisdom to be doubly defective.

In the first place, the appearance of a singularity at $r = \pi R/2$ is only an illusion, for the de Sitter coordinates cover only a portion of a larger spacetime that, by any reasonable standards, is singularity free. An understanding of this point emerged from the papers of Felix Klein (1918a, b) and Cornelius Lanczos (1922b).[12] A detailed discussion of de Sitter spacetime was also given in Arthur Stanley Eddington's influential book *The Mathematical Theory of*

Relativity (1923). Although Eddington clarified the nature of the de Sitter "singularity," the way he presented the matter revealed an imperfect understanding of the Schwarzschild solution.

Eddington rewrote the line element of the de Sitter metric in the form[13]

$$ds^2 = (1 - r^2/R^2)^{-1}\, dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) - (1 - r^2/R^2)\, dt^2 \quad (1.6)$$

He noted that in (1.6) there is a "singularity" at $r = R$ "similar to the singularity at $r = 2M[\alpha]$ in the solution for a particle of matter," a reference to the Droste–Schwarzschild line element (1.2). Comparison of (1.6) with (1.2) raised the question of whether the de Sitter spacetime is really empty of matter.

> Must we not suppose that the former singularity [in (1.6)] also indicates matter—a "mass-horizon" or ring of peripheral matter necessary in order to distend the empty region within. If so, it would seem that de Sitter's world cannot exist without large quantities of matter . . . he has merely swept the dust away into unobserved corners. (Eddington 1923, p. 165)

This is undoubtedly a reference to Einstein's critique of the de Sitter solution, which had ended with the assertion that in the de Sitter world matter is "entirely concentrated in the [singular] surface $r = \pi R/2$" (Einstein 1918, p. 272). However, Eddington's answer to his rhetorical question was an emphatic no. "A singularity in ds^2 does not necessarily indicate material particles, for we can introduce or remove such singularities by making transformations of coordinates" (ibid., p. 165). He continued:

> The whole of de Sitter's world can be reached by a process of continuation; that is to say the finite experiences of an observer A extends only over a certain lune; he must then hand over the description to B whose experience is partly overlapping and partly new; and so on by overlapping lunes. . . . [In this way] we arrive at de Sitter's complete world without encountering any barrier or mass-horizon. (ibid., p. 166)

Here then we have a codification of the first small success in understanding the difference between genuine singularities and coordinate singularities.

Knowledge of the nature of the $r = \alpha$ Schwarzschild singularity was harder won. Lanczos (1922a) hinted at the pseudo-nature of this singularity. Two years later in a letter to *Nature*, Eddington (1924) employed the transformation

$$t' = t - \alpha \ln(r - \alpha) \quad (1.7)$$

which recast (1.2) in the form

$$ds^2 = dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) - dt'^2 + (\alpha/r)(dr - dt')^2 \quad (1.8)$$

(Actually, there is a missing factor of 2 in Eddington's formula which, in the units he was using, contained the term $2M \ln(r - M)$ rather than $2M \ln(r - 2M)$.) The blowup of $g_{rr} = (1 - \alpha/r)^{-1}$ in (1.2) at $r = \alpha$ has been replaced by tame behavior of $g'_{rr} = (1 + \alpha/r)$ in (1.8). In addition, in the new coordinates the determinant of the metric potentials is non-vanishing at $r = \alpha$ and, therefore, the contravariant metric components have finite values. It is on this basis that Eddington is sometimes credited with showing that the Schwarzschild metric is non-singular at $r = \alpha$ (see Misner, Thorne, and Wheeler 1973, Box 31.1). However, in Eddington's coordinates $g'_{44} = g_{44}$, which vanishes at $r = \alpha$. Hindsight wisdom is needed to see that the vanishing of g_{44} and g'_{44} is an indication not of a breakdown in the metric but of the fact that $\partial/\partial t'$ and $\partial/\partial t$ turn from timelike to null at the Schwarzschild radius. In any case, the purpose of Eddington's communication was to compare Einstein's and Whitehead's theories of gravity, and Eddington himself made no claim to have clarified the status of the Schwarzschild radius.[14]

The first explicit and self-conscious demonstration that the $r = \alpha$ Schwarzschild singularity was merely a coordinate singularity was due to Georges Lemaître (1932). Here was a golden opportunity to illustrate within the same solution the difference between genuine and coordinate singularities and to provide a rough criterion for telling the two apart. The Kretschmann curvature invariant $K = R_{abcd}R^{abcd}$ and other curvature scalars as well blow up as $r = 0$ is approached ($K \sim 1/r^6$), whereas they remain well behaved as $r = \alpha$ is approached. As we will see in chapter 2, trying to separate coordinate singularities from genuine singularities on this basis is too crude, since genuine singularities are not always signaled by the ill behavior of curvature invariants. But even such an imperfect criterion would have had a salutary effect on the discussions of these matters.

In the actual case, however, even this rudimentary understanding continued to elude some of the most able researchers. An especially glaring example is Einstein's 1939 treatment of the Schwarzschild radius. Einstein chose to work in isotropic coordinates in which the Schwarzschild line element (1.2) takes the form

$$ds^2 = (1 + \mu/2r)^4(dr^2 + r^2\, d\theta^2 + r^2 \sin^2\theta\, d\phi^2) - \left(\frac{1 - \mu/2r}{1 + \mu/2r}\right)^2 dt^2 \quad (1.9)$$

where $\mu = 2\alpha$. Einstein noted that at $r = \mu/2$ a clock would "go at the rate zero" and that light rays and material particles starting from $r > \mu/2$ would "take an infinitely long time (as measured in 'coordinate time') in order to reach the point $r = \mu/2$." He concluded that "In this sense the sphere $r = \mu/2$ constitutes a place where the field is singular" (Einstein 1939, p. 922).

The aim of Einstein's article was to show that the "'Schwarzschild singularities' do not exist in physical reality" (ibid., p. 936). The basis of this conclusion was the claim that a gravitational field with such singularities

cannot arise as the exterior field of a realistic system of masses. Other research workers had put forward a weaker version of this claim by noting that the Schwarzschild radius for any known astronomical body or elementary particle is well within the body itself where, of course, the exterior Einstein field equations do not apply. This "pragmatic attitude," as Eisenstaedt (1989) has called it, towards the Schwarzschild singularity was maintained at least as late as 1960 (see, for example, Synge 1960, p. 183). Towards the end of demonstrating a stronger, in-principle version of this claim, Einstein considered a spherical cluster of particles, each of which moves in a circular orbit. He found that, in a condition of static equilibrium, as the radius of the cluster approaches the Schwarzschild radius, the particles would eventually have to move faster than light. In hindsight, what Einstein's calculations showed was not that a massive body cannot collapse inside its Schwarzschild radius but that if it contracts sufficiently near to this radius it cannot remain in static equilibrium. And we now know that the further contraction which the body cannot resist may very well eventuate in a genuine curvature singularity (see Thorne 1994, Ch. 3).

It is obvious that in 1939 Einstein was working in ignorance of Lemaître's 1932 article, which was first published in an obscure Belgian journal and then republished in a less obscure but not widely read Belgian journal. Nevertheless, it is surprising that in 1939 Einstein failed to address the question of whether the $r = \mu/2$ "singularity" could be removed by a coordinate transformation and simply concluded that "the field is singular" there. Could it be that Einstein was applying his 1918 definition of singularity and making the same mistake that Hilbert made in 1917? This is unlikely since in his critique of de Sitter's solution Einstein qualified his conclusion that a singularity exists at $r = \pi R/2$ with an "until further notice" clause, an unnecessary concession if Hilbert's restriction on the nature of the coordinate transformation needed to remove the singularity had been in effect. A much more likely explanation of how Einstein misled himself is contained in the above quoted passage about the need for an infinite amount of coordinate time for a particle to reach $r = \mu/2$ from $r > \mu/2$. The irony, of course, is that the fact Einstein cites is an indication of the breakdown of the t coordinate rather than a proof of the physical inaccessibility of $r = \mu/2$. A second irony derives from the observation that in the original dispute which gave rise to Einstein's attempt to define the notion of spacetime singularity, de Sitter defended his metric by appealing to the physical inaccessibility of the $r = \pi R/2$ singularity. Here we witness a potentially confusing entanglement of the problem of singularities and the problem of horizons in cosmology. I will return to the latter topic in the next section.

To return to Einstein's attempt to characterize spacetime singularities, the second facet of his definition, which sought to capture the notion of "at a finite distance," is also defective. In the de Sitter spacetime the supposed singular region can be reached by spacelike curves of finite proper length. But in other cases—for example, the standard big bang models—the singularity

can only be reached by timelike or lightlike curves. The proper length of a lightlike curve is, of course, 0, and if arbitrary timelike curves are permitted, then what would be regarded as future or past infinity can be reached by a timelike curve of finite proper length. Current attempts to capture the notion of "at a finite distance" still use Einstein's idea of a curve which, in some appropriate sense, has finite length (see chapter 2). But there is a subtle difference. Modern constructions use a *half*-curve, a curve which starts at some point p_0 (Einstein's origin point) and which is extended as far as possible in some direction from that point. For such a half-curve, there is no endpoint corresponding to Einstein's p—the singularity does not have a location in spacetime. Although this may seem to be nitpicking, it is of crucial importance for the interpretation of singularities and, as will be seen in the following section, the misunderstanding of this point had an adverse affect on Einstein's attempt to deal with the problem of motion of a particle in a gravitational field.

Leaving aside the technical niceties of Einstein's attempt to define spacetime singularities, it is noteworthy that most of the subsequent attempts have shared Einstein's philosophy on one crucial point; namely, if something nasty happens to the spacetime metric, but the nastiness happens only "at infinity," then no singularity in the spacetime is indicated. However, as Roger Penrose's discussion of "naked singularities" and "cosmic censorship" has revealed, nastiness at infinity can still be disruptive to physics, and if it is disruptive enough it deserves the name singularity (see chapters 2 and 3).

Since the reader who is new to these issues may be reeling from trying to follow the twists and turns of the discussion, it may be well to step back from the complexities in order to gain more perspective on Einstein's attitude towards singularities.

1.3 Einstein's intolerance of singularities

The visceral dislike of singularities which Einstein displayed in 1917–18 remained with him until the end of his life. Here is Peter Bergmann, who served as a research assistant to Einstein at the Institute for Advanced Study, speaking of Einstein's attitude towards singularities:

> It seems that Einstein always was of the opinion that singularities in classical field theory are intolerable. They are intolerable from the point of view of classical field theory because a singular region represents a breakdown of the postulated laws of nature. I think one can turn this argument around and say that a theory that involves singularities and involves them unavoidably, moreover, carries within itself the seeds of its own destruction. . . . (Bergmann 1980, p. 186)

In the first years following the formulation of the GTR, Einstein's attitude towards singularities may have wavered, but Einstein's published papers and

the correspondence of which I am aware indicate that Bergmann has captured Einstein's mature attitude. This conclusion raises three questions.

First, how is Einstein's rejection of singularities to be reconciled with his willingness, in connection with the problem of motion, to treat massive particles as singularities in the gravitational field? Second, how did Einstein propose to resolve the tension between his intolerance of singularities and the existence of singularities in solutions to his gravitational field equations? Third, what exactly was the basis of his rejection of singularities? And apart from understanding Einstein's attitude, this book is dedicated to understanding whether his horror of singularities is justified.

To take first questions first, Einstein wanted to show that the postulate that the world line of a massive neutral test particle is a timelike geodesic need not be taken as a separate axiom of GTR but is in fact a consequence of the gravitational field equations, and several publications stretching over twenty years are devoted to this goal (see Einstein and Grommer 1927; Einstein 1927; Einstein, Infeld, and Hoffman 1938; and Einstein and Infeld 1949). Towards this end he was willing to treat the particle as a singularity in the gravitational field. His attitude towards this procedure is summarized in a letter to Ludwig Silberstein dated 18 February 1941:

> Of course, in a complete field theory the positing of singularities is altogether forbidden. In the present case the introduction of singularities is justified because it allows the treatment on the basis of the gravitational field alone of a problem of which "matter" is a part, without having to use a theory of the latter. (EA 21-090)[15]

It is irresistible to see a strong analogy between Einstein's procedure here and his path to the special theory of relativity (STR). H. A. Lorentz's version of the special principle of relativity was based on what Einstein called a "constructive theory" (Einstein 1954, p. 228)—in this instance, an elaborate theory of the electromagnetic constitution of matter which made predictions about how matter behaves when it moves through the ether. By contrast, Einstein's approach employed a "principle-theory" which required no commitments about the constitution of matter and only minimal commitments about the behavior of light and electromagnetic waves. In parallel, Einstein's strategy on the problem of motion in a gravitational field was to avoid questions about the nature of matter by treating a test particle as a pole singularity in a solution of the matter-free field equations $G_{ab} = 0$ (or equivalently, $R_{ab} = 0$) for $\Lambda = 0$.[16]

There are several perplexing features of Einstein's treatment of the problem of motion, but the most perplexing of all lies in the fact that singularities in the spacetime metric cannot be regarded as taking place at points of the spacetime manifold M.[17] Thus, to speak of singularities in g_{ab} as geodesics of the spacetime is to speak in oxymorons. One might hope to give some sense to this talk by representing singularities as ideal points adjoined to M. But there are several different ways to accomplish the

representation, and they yield inequivalent characterizations of singularities. Moreover, all of the extant procedures yield unexpected and counterintuitive results, e.g., the singular points may not be Hausdorff separated from the regular points of M (see chapter 2).

Treating spacetime singularities as lying in the spacetime is not only mathematically unjustified but is apparently also inconsistent with Einstein's own attitude on the ontological status of spacetime.[18] During 1913–14 when he was searching for the relativistic gravitation field equations Einstein used the notorious "hole argument" to justify rejecting the requirement of general covariance (see Norton 1987). After reaffirming general covariance he proposed to escape the hole argument by maintaining that

> If we imagine the gravitational field, i.e., the functions g_{ik}, to be removed, there does not remain a space of type (1) [Minkowski spacetime], but absolutely *nothing*, and also no "topological space." For the functions g_{ik} describe not only the field, but at the same time also the topological and metrical properties of the manifold. ... There is no such thing as empty space, i.e., a space without field. Space-time does not claim an existence of its own, but only as a structural quality of the field. (Einstein 1961, p. 155)

Einstein scholars have not provided an explanation of how Einstein was able to reconcile the apparent conflict between the view quoted and his treatment of the geodesic postulate in terms of singular regions of spacetime.

There might still seem to be room for compromise here. Start with a non-singular background spacetime M, g_{ab}. Following Einstein, let us avoid making any assumptions about the constitution of matter by treating a massive particle as a point object with world line $z^a = z^a(\tau)$ parameterized by proper time τ. The associated energy–momentum tensor can be postulated to be $T^{ab}(x) = m \int \dot{z}^a \dot{z}^b \delta(x - z(\tau))\, d\tau$ where $\dot{z}^a = dz^a/d\tau$ and m is the mass of the particle. The conservation law $\nabla_a T^{ab} = 0$ can then be used to show that $z^a(\tau)$ is a geodesic of M, g_{ab}.[19] But such a demonstration does not constitute a derivation of the geodesic postulate from the field equations $G_{ab} = 8\pi T_{ab}$. These equations do entail the conservation law, but for the singular T^{ab} postulated here the field equations are not meaningful in even a distributional sense. The widest class of spacetime metrics for which the Riemann tensor makes sense as a distribution is arguably the *regular metrics* as defined by Geroch and Traschen (1987) (see chapter 2); but this class is not wide enough to encompass sources concentrated on one-dimensional submanifolds.

There are alternative approaches to deriving the geodesic postulate that do not involve the problematics of singularities and are relatively neutral about the constitution of matter. For example, Fock (1939) showed that under mild assumptions about a non-singular T^{ab}, the conservation law $\nabla_a T^{ab} = 0$ entails the geodesic hypothesis for sufficiently small bodies whose self-gravity is weak.[20] Einstein evinced no interest in such approaches that made any assumptions about the nature of matter. His final official word on the subject is found in a joint paper with Infeld:

> All attempts to represent matter by an energy–momentum tensor are unsatisfactory and we wish to free our theory from any particular choice of such a tensor. Therefore we shall deal here only with gravitational equations in empty space, and matter will be represented by singularities of the gravitational field. (Einstein and Infeld 1949, p. 209)

I will return below to Einstein's struggles with the problem of motion and singularities, but for the moment I turn to the tension between the view that "in a complete theory the positing of singularities is altogether forbidden" and the existence of singularities in solutions to EFE. There are at least three ways to resolve the tension, which will be discussed in turn.

The first resolution is to argue that singularities are artifacts of the idealizations of the models in which they occur. This is the tack Einstein took in 1931 in reaction to the big bang singularity of the Friedmann model:

> Here one can try to get out of the difficulty by pointing out that the inhomogeneity of the stellar matter makes illusory our approximate treatment. (Einstein 1931, p. 237)

Eight years later Einstein took a similar line with respect to what he regarded as the singular aspect of the Schwarzschild solution. As we saw above in section 1.2 he argued that it is not possible to build up a field containing such "singularities" (as at $r = \mu/2$ in (1.9)) from physically realistic arrangements of gravitating matter (Einstein 1939). The work of Oppenheimer and Volkof (1939) and Oppenheimer and Snyder (1939) indicated that the gravitational collapse of a star could produce a singularity. There is evidence that as late as 1942 Einstein was either not aware of this work or else brushed it aside.[21] In the latter case his justification would no doubt have been that the singularity predicted was an artifact of the assumed perfect spherical symmetry.

In the end, however, this first resolution will not suffice. Einstein himself produced evidence that singularities are not isolated features of solutions to his gravitational field equations. In 1941 he argued that there are no solutions to the vacuum field equations which are associated with a positive mass, which are static and asymptotically Minkowskian, and which are singularity free (Einstein 1941). Related results had already appeared in the literature (see Lichnerowicz 1939). In 1943 Einstein and Pauli generalized Einstein's 1941 result to cover gravitational theories of the five-dimensional Kaluza type (Einstein and Pauli 1943). However, the fact that GTR and related theories seemed incapable of producing particle-like solutions without singularities[22] did not endear Einstein to singularities but served rather to make him dissatisfied with these theories. Had Einstein lived until the end of the 1960s he would no doubt have agreed that the theorems of Hawking and Penrose show unequivocally that spacetime singularities in GTR are not artifacts of specialized models. But these theorems would most likely have served to increase his dissatisfaction with GTR.

The second resolution is to appeal to the cosmological constant term in the EFE. Λ is not a completely effective holy cross against the vampire of

singularities, but it can help. For example, a large enough positive value for Λ will ward off the big bang singularities in the Friedmann models. This would not have been an attractive alternative to the Einstein who reportedly said that the introduction of Λ was the "biggest blunder he ever made in his life" (see Gamov 1970, p. 44). However, Einstein's attitude towards Λ was more complicated than this quotation indicates. In an appendix "On the 'Cosmologic Problem'" added to the second edition of *The Meaning of Relativity*, Einstein opined that the "cosmologic member" (i.e., the Λg_{ab} term in the gravitational field equations) "is to be rejected from the point of view of logical economy" (Einstein 1955, fn pp. 120-121). His reasoning is explained in a footnote:

> If Hubble's expansion had been discovered at the time of the creation of the general theory of relativity, the cosmologic member would never have been introduced. It seems now so much less justified to introduce such a member into the field equations, since its introduction loses its sole original justification—that of leading to a natural solution of the cosmologic problem. (Einstein 1955, fn p. 121)

The "cosmologic problem" to which Einstein refers is to reconcile a finite average matter density with the gravitational field equations, a problem Friedmann resolved by allowing for the expansion of the universe. The quoted passage is consistent with the attitude that *other* cosmological problems could justify the reintroduction of a Λ term. One such problem was already mentioned in Einstein's appendix: the Friedmann model in conjunction with the then existing astronomical measurements of the Hubble constant led to an age for the universe that was too short.[23] But Einstein showed no inclination to use Λ to solve this problem. Nor is there any mention of Λ as a solution to the problem of singularities.

The third resolution is to say that the field equations (without cosmological constant term) break down under the conditions that singularities are predicted to occur. In the same appendix "On the 'Cosmologic Problem,'" Einstein expressed his doubts about the consequence of the spatially closed Friedmann models that "for the time of the beginning of the expansion the metric becomes singular and the density . . . becomes infinite."

> For large densities of field and matter, the field equations and even the field variables which enter into them have no real significance. One may not therefore assume the validity of the equations for very high density of field and matter, and one may not conclude that the "beginning of expansion" must mean a singularity in the mathematical sense. (Einstein 1955, p. 123)

Out of context this quotation seems to represent an uncharacteristic loss of nerve on Einstein's part: if we refuse to trust the most fundamental equations of physics to work under extreme conditions, then we will be barred from making many interesting predictions. And if the EFE do break down under extreme conditions, what will take their place? Here Einstein would seem to

face a dilemma. The EFE can be replaced either by some other classical field equations or by some system of quantum mechanical equations. In the former case one can wonder why the new equations can be trusted anymore than the EFE to work under extreme conditions. In the latter case one can imagine that under extreme densities quantum effects emerge and somehow prevent the formation of singularities. But Einstein would not have seen this escape from singularities as a live option, for he never gave up the hope of deriving quantum constraints from classical field theory.

Placed in proper context the quotation does not signal any loss of nerve, for Einstein had long been dissatisfied that his field equation had a T_{ab} term on the right-hand side.[24] And Einstein's work on unified field theory can be seen as a seizing of the first horn of the apparent dilemma. His program of unification focused on the gravitational and electromagnetic fields, with the primary goal of treating both as aspects of the spacetime geometry. One consequence of the unification he expected was that matter could be treated as a derived rather than a basic concept in that the unified field equations, with no matter source term, would admit singularity free, particle-like solutions. And finally, quantum constraints were also to be derived, perhaps by conditions overdetermining the classical field variables.[25] With this as background, we can understand the following optimistic passage from *The Meaning of Relativity*:

> The present relativistic theory of gravitation is based on a separation of the concepts of "gravitational field" and of "matter". It may be plausible that the theory is for this reason inadequate for very high density of matter. It may well be the case that for a unified theory there would arise no singularity. (Einstein 1955, fn p. 124)[26]

Had Einstein become as pessimistic as most of his contemporaries about the prospects of a unified field theory, he would have been forced to conclude that the third option for dealing with singularities, as well as the first two, was closed off. This closing off would have necessitated a reassessment of his rejection of spacetime singularities.[27] Such a reassessment was suggested by the man who spent the early part of his career trying to establish the experimental basis of Einstein's GTR, Erwin Freundlich, now named Finlay-Freundlich. In 1951 Finlay-Freundlich's monograph on *Cosmology* appeared as Vol. 1, No. 8 of the Foundations of the Unity of Science series. In a section entitled "General Remarks concerning Singularities in the Cosmological Problem," he commented on the big bang singularity in the Friedmann models. He noted that an initial singularity can be avoided by introducing a large positive value for the cosmological constant, but termed the introduction "a serious step" for which there "appears otherwise no necessity." He continued:

> We must ask therefore, since singularities arise when we get excessively large values of the [mass] density . . . , whether our fear of excessive densities, the opposite of *horror vacui*, is justified and not perhaps the last reminder of a subconscious yearning for a harmonious universe. (Finlay-Freundlich 1951, p. 49).

There is no indication that Einstein had read Finlay-Freundlich's *Cosmology* or paid attention to his suggestion any more than that he paid much heed to the opinions of any of his contemporaries on these matters.[28] Only from Einstein's own writings can we hope to discern the reasons for his aversion to spacetime singularities. Early on (circa 1918) Einstein seems to have thought that spacetime singularities involve singularities in the matter sources and that the latter singularities are absurd. But it later became evident that spacetime singularities need not hide mass concentrations, singular or otherwise.

A good starting point for exploring the mature Einstein's aversion to spacetime singularities is a joint paper by Einstein and Nathan Rosen published in *The Physical Review* in 1935. They wrote:

> A singularity brings so much arbitrariness into the theory . . . that it actually nullifies its laws. . . . Every field theory, in our opinion, must therefore adhere to the fundamental principle that singularities of the field are to be excluded. (Einstein and Rosen 1935, p. 73)

Taken by itself this quotation is potentially misleading. What Einstein and Rosen were objecting to was not so much spacetime singularities in general as the idea that material particles can be treated as singularities of a gravitational field, an idea that, as noted above, was explored by Einstein himself from time to time. In the paper in question this idea was rejected by Einstein and Rosen for the reason that such a singularity "nullifies" the laws of the theory. They explained:

> A pretty confirmation of this was imparted in a letter to one of the authors by L. Silberstein. As is well known, Levi-Civita and Weyl have given a general method for finding axially symmetric static solutions of the gravitational equations. By this method one can readily obtain a solution which, except for two point singularities lying on the axis of symmetry, is everywhere regular and is Euclidean at infinity. Hence if one admitted singularities as representing particles one would have here a case of two particles not accelerated by their gravitational action, which would certainly be excluded physically. (ibid., p. 73)

To understand the content of this passage requires a digression.

What Silberstein had communicated to Einstein was a rediscovery of a version of the Curzon (1924a, b) bipolar solution, which belongs to the Weyl class (1917, 1919) of static axially symmetric solutions. If the above quotation accurately describes the solution and if singularities in the solution are taken to represent massive particles, then as Silberstein went on to claim, not every solution to the vacuum EFE $G_{ab} = 0$ (or $R_{ab} = 0$) would be physically admissible. Einstein and Rosen seemed to agree, for the main point of their paper was to contemplate a modification of the EFE which would admit of

non-singular solutions for the static spherically symmetric case. A particle would then be represented not by a singularity but by the multiple connectedness of the space in the form of a "bridge" joining two identical sheets of space.

Encouraged by the remarks in the Einstein–Rosen paper, Silberstein went on to publish his calculations in the February 1, 1936 issue of *The Physical Review*.[29] He concluded:

> Thus our solution ... corresponds to a complete absence of matter ... everywhere, except only at the points A, B themselves, the "mass centers." There are thus ... no stresses between A and B to keep them apart like a stiff rod. And yet (the solution being stationary, invariable in time) the two points are fixed relatively to each other instead of falling towards each other, which flagrantly contradicts man's most ancient, primitive experience.
>
> Here, then, is an example of a perfectly rigorous solution to Einstein's field equations which does not correspond to reality. If, therefore, these equations are to be retained, one cannot consider "material particles" (mass points) as singularities of the field. (Silberstein 1936, p. 270)

Although Einstein had agreed to the gist of this conclusion, its appearance in print prompted him to examine Silberstein's calculations more closely. Together with Rosen he published a letter in *The Physical Review* rebutting Silberstein's claims.

> In a recent paper, Silberstein attempts to show the incorrectness of the general theory of relativity. His reasoning is as follows:
>
> (a) I set up a static solution of the gravitational field equations which has two singular points and which is everywhere else free of singularities.
>
> (b) The two particles so represented are not accelerated in each other's gravitational field, in contradiction with experience. Hence the gravitational field equations of the general relativity theory are incorrect.
>
> We should like to point out the following. Even if (a) were the case the conclusion (b) would not be justified. For in a field theory only a representation of masses which is free from singularities can be accepted, since at a singularity the laws of the field are violated. However the assertion (a) is not correct. We shall show that the solution given by Silberstein has singularities outside the two points. (Einstein and Rosen 1936, p. 404)

Foisting (b) on Silberstein was a little disingenuous. Silberstein was not claiming to have shown that EFE are incorrect but only that, assuming (a), *either* the field equations are incorrect *or else* "one cannot consider 'material particles' (mass points) as [pole] singularities in the field," a claim Einstein would not have disputed. The real bone of contention was (a). Silberstein assumed, and Einstein accepted, that the singularities in the Curzon–Silberstein solution are simple pole singularities. The first intimation that this assumption was incorrect did not come until thirty years later when Gautreau and Anderson (1967) showed that the singularities seem to have an angular character—whether the Kretschmann curvature scalar blows up depends on

the direction from which the singular 'points' are approached. Stachel (1968) clarified the matter by showing that the singular 'points' are not pointlike at all but have a non-trivial topological structure. The local and bizarrely complicated global structure of the singularity in the Curzon monopolar solution has only recently received its definitive elucidation by Scott and Szekeres (1986a, b).

The main disagreement between Einstein and Silberstein was whether, as Silberstein claimed, the axis joining the two singular 'points' is non-singular. For the Curzon bipolar solution Silberstein's claim is incorrect. But seeing how the claim is incorrect helps to reveal just how subtle is the business of singularities. In Weyl's coordinates the line element takes the form

$$ds^2 = e^{-2U}[e^{2k}(d\rho^2 + dz^2) + \rho^2\, d\phi^2] - e^{2U}\, dt^2 \qquad (1.10)$$

where U and k may be functions of ρ and z. (Here the z-axis is the axis of symmetry, ρ is the radial coordinate, and ϕ is the angular coordinate.) In order that the axis joining the two singular 'points' be non-singular, it is necessary that $\lim_{\rho \to 0} k = 0$, otherwise the ratio of the circumference to the radius of an infinitesimal circle enclosing the axis would not be 2π. Einstein and Rosen (1936) noted that this condition fails for the Curzon–Silberstein bipolar metric.[30] And yet there is also a sense in which the axis is non-singular. If the components of the curvature tensor are computed in a suitable frame along a geodesic approaching the axis, they remain well behaved—there is no blowup or wild oscillation. What we have here is a non-curvature singularity or in modern parlance a quasi-regular singularity (see chapter 2). The intuitions shared by Einstein and Silberstein in discussing the implications of spacetime singularities simply do not begin to do justice to the surprising complexities that spacetime singularities can display.

It is evident from the controversy with Silberstein that Einstein's attitude towards singularities was shaped in part by the problem of motion. But Einstein's published response to Silberstein reveals a reason why he thought that, independently of the problem of motion, singularities must be excluded: "in a field theory only a representation of masses which is free of singularities can be accepted, since at a singularity the laws of the field are violated." Einstein is surely right that, whatever the technical details of a definition of spacetime singularities, it should follow that physical laws, in so far as they presuppose space and time, are violated or, perhaps more accurately, do not make sense at singularities. This is a good reason for holding (as was urged above) that singularities are not part of spacetime. But why should singularities be seen as spelling such a disaster for physics that any spacetime theory involving singular solutions must be deemed inadequate? Consider, for example, the Friedmann–Robertson–Walker (FRW) models of the big bang. There is a rigorous sense in which these models exhibit Laplacian determinism, by which the physical state of the universe at any time uniquely fixes the state at any other time. Within these models the application of determinism

permits us to retrodict the existence of a big bang singularity and, in spatially closed models with sufficient mass, to predict a big crunch singularity. Thus, the application of the EFE implies that time is finite in the past, and, if there is sufficient mass, also in the future; but within this finite stretch physics is conducted as usual. There may be various psychological objections to the finiteness of time, but it is far from obvious why physical laws that lead to this conception thereby impugn themselves. Perhaps it will be said that these laws are defective because they entail the existence of a big bang and a big crunch without telling us what happened before the bang or after the crunch. But this objection misses the point. As Einstein said, physical laws break down at spacetime singularities, and for the big bang and the big crunch they break down so strongly that it is physically meaningless to talk about the 'before' and the 'after' (see chapters 2 and 7).

It remains to contemplate the possibility that singularities might occur not at the beginning or end of time but, so to speak, in the interior of spacetime where they could disrupt physics. An attempt to distinguish between disruptive and non-disruptive singularities might be seen to be behind de Sitter's response to Einstein's (1918) critique of the de Sitter solution. De Sitter conceded—unnecessarily as we now know—that his solution violated Einstein's stricture that in a physically admissible spacetime there are no discontinuities at finite distances. He continued:

> This postulate, however, in the form in which it is enounced by Einstein, is a *philosophical*, or metaphysical, postulate. To make it a *physical* one, the words "all points *at finite distances*" must be replaced by "all *physically accessible* points". And if the postulate is thus formulated, my solution ... does fulfill it. For the discontinuity arises for $r = r_1 = \pi R/2$.
>
> This is at a finite distance in space, but it is physically inaccessible. ... The time needed for a light ray, and a fortiori by a moving material point, to travel from any point r, ψ, ϕ to a point r_1, ψ_1, ϕ_1 (ψ_1 and ϕ_1 being arbitrary) is infinite. [And by symmetry, the time needed to travel in the opposite direction is also infinite.] The singularity at r_1 can thus never affect any physical experiment .... (de Sitter 1918, p. 1309)

There are several things here that need to be disentangled. The class of non-disruptive singularities in the sense of singularities that do not interfere with determinism and predictability is not coextensive with the class of singularities that are physically inaccessible—as already noted, the big bang singularity is certainly accessible but it is not disruptive. The disruptive vs. non-disruptive distinction is today pursued under the labels of naked vs. non-naked singularities (see chapter 3). De Sitter's notion of physically inaccessible singularities does call to mind one way in which a singularity can be non-naked; namely, by being hidden inside a black hole. But the most obvious application of the notion of physical inaccessibility to de Sitter spacetime calls attention not to black holes but to two other features of general relativistic cosmological models—particle horizons and event horizons (see chapter 6). Because de Sitter spacetime possesses both features there are regions that have been and forever will be causally disconnected from one another. The existence of particle horizons in the standard big bang models makes it difficult for these models to satisfactorily explain the isotropy of the cosmic microwave radiation (or so it has been claimed), and this "horizon problem" has spawned inflationary cosmology and a number of other competing models (see chapter 5). Once again we have an example where the pioneers of GTR hit upon the problems that are now at the heart of gravitational research but did not possess the analytical tools to properly investigate the problems. Nor was there any call to develop those tools until it became clear that singularities could not be ignored.

1.4 Acausality and time travel

By contrast with spacetime singularities, the recognition of the fact that solutions to EFE can involve acausal spacetime structures did not come until comparatively late.[31] Van Stockum (1937) found an exact solution to EFE for a source consisting of an infinite rotating cylinder of dust; but it was not realized until decades later that the exterior spacetime contains closed timelike curves (see Tipler 1974). The first solution known to have this feature was published by Gödel (1949a). And while singularities have been a focal point of research in relativistic gravitational theory for the past thirty years, it is only quite recently that closed timelike curves and the like have received much attention in the physics literature (see chapter 6).

Gödel (1949b) contributed a brief account of his discovery to the Schilpp (1949) volume, *Albert Einstein: Philosopher–Scientist*, in which he attempted to use his technical results to support the idealistic conclusion that time is not real (see the Appendix to chapter 6). In his reply to criticisms, Einstein (1949b) brushed aside the relation of relativity theory to idealism in order to concentrate on what he took to be the fundamental problem raised by Gödel's solution. He invited the reader to consider three points A, P, B lying on a timelike world line, with P between A and B. Does it make sense, he asked, to assign an arrow to the world line, indicating that, say, B is before P and A after P? A positive answer is warranted, Einstein asserted, if it is possible to send a signal from B through P to A but not vice versa. Einstein continued:

> If, therefore, B and A are two, sufficiently neighboring, world-points, which can be connected by a time-like line, then the assertion: "B is before A," makes physical sense. But does this assertion make sense, if the points, which are connectable by the time-like line, are arbitrarily far separated from each other? Certainly not, if there exist point-series connectable by time-like lines in such a way that each point precedes temporally the preceding one, *and if the series is closed in itself*. In that case the distinction "earlier–later" is abandoned for world points which lie far apart in a cosmological sense, and

> those paradoxes, regarding the *direction* of causal connection, arise, of which Mr. Gödel has spoken. (Einstein 1949b, p. 688)

I read Einstein as providing, if somewhat imperfectly, a distinction between two features of time: *time directionality* and *time order*. The spacetime of Gödel's solution admits of a globally consistent assignment of time directionality or distinction between past and future at each event, but it does not admit of a consistent assignment of time order to events. The point is not quite captured, as the quotation from Einstein would suggest, by saying that the distinction 'earlier–later' makes sense for nearby events but has to be abandoned for events that are far apart. But such niceties aside, what are the paradoxes, "regarding the *direction* of causal connection," that arise from Gödel's solution? Again there is a slight mischaracterization of the problem in the quotation: there is no ambiguity in the direction of causation in the Gödel model—causal propagation is assumed to always take place in the future direction which, as already noted, is a globally well-defined notion in this model. The problem is rather that the kinds of causal stories we are used to telling threaten to degenerate into gibberish in the Gödel universe. For Gödel's solution seems to allow for the physical possibility of a journey into one's own past. Once one is there, why couldn't one then proceed to undo what has already been done? Of course, it is nonsensical to suggest that one can change what has already happened, but what is to prevent the deeds that would lead to the absurdity?

Einstein's concern with such paradoxes was expressed in his closing remarks.

> Such cosmological solutions [that allow time travel] of the gravitational equations (with non-vanishing Λ-constant) have been found by Mr. Gödel. It will be interesting to weigh whether these are not to be excluded on physical grounds. (Einstein 1949b, p. 688)

In the first sentence of the quotation lies an implicit criticism of the Gödel solutions: they require a non-vanishing cosmological constant. Given Einstein's rejection of this term in his field equations, there was reason for hope that acausal solutions could be ignored on the grounds that they were confined to the $\Lambda \neq 0$ cases. Today we know that this hope is not realized. Thus we must face the question raised in the second sentence of the quotation: Are such solutions to be "excluded on physical grounds"?

1.5 Singularities and acausalities together

To return to the analogy offered in section 1.1, spacetime singularities and acausal spacetime structures pose challenges for GTR not unlike the challenge the measurement problem poses for QM. Are the nastier sorts of singularities and acausalities to be excluded on physical grounds? If so, can the exclusion

be enforced on some non-ad hoc basis within GTR or some natural extension of the theory, or must the theory be modified in some substantial way in order to achieve the exclusion? If on the other hand we elect to try to find ways to peacefully coexist with singularities and closed time loops, what price does the coexistence exact of our concepts of physical laws, causality, and free will?

These are certainly questions worthy of study. But why study the problems of singularities and acausality together in one work? The short answer, which can only be justified by a lengthy analysis, is that they are not separable problems. The theorems of Hawking and Penrose that prove the existence of singularities in wide classes of solutions to EFE assume the absence of closed time loops. This naturally raises the question of whether acausal features of spacetime can prevent the occurrence of singularities. In the other direction, attempts to prove that there are obstacles to operating a time machine which would create closed time loops assume that spacetime singularities are not formed through the operation of the machine. This raises questions about what kinds of singularities might be tolerated by a time machine operator. And third, Penrose's cosmic censorship hypothesis, which seeks to show that the GTR does not permit the formation of naked singularities under physically reasonable conditions, would count some forms of acausality as violating cosmic censorship.

The first order of business is to try to say more precisely what we are talking about. Chapter 2 discusses various analyses of the concept of spacetime singularity and the means of classifying the types of singularities that can occur in relativistic spacetimes. Because of the technical nature of the material, some readers will find chapter 2 heavy going. If you are one of these, gentle reader, do not lose heart, for the subsequent chapters are designed to stand largely on their own. Thus, you may want to skim chapter 2 and then turn to the topics that interest you the most, while being prepared to occasionally return to chapter 2 to clarify some point.

The topics of the subsequent chapters were chosen partly to illuminate the issue of singularities in GTR and partly to emphasize the connection of these issues to philosophical concerns about predictability and determinism, paradoxes of infinity, the common cause principle and the nature of scientific explanation, backwards causation and time travel, etc. Once the connections are revealed, it is hard to see how philosophers of science can continue to write on these topics without taking into account the implications of GTR.

Chapter 3 reviews various formulations of the cosmic censorship hypothesis and weighs the evidence that has been amassed for and against censorship. It also tackles the difficult question of how much a disaster for physics it would be if cosmic censorship should fail. One startling consequence of the failure is discussed in chapter 4. If singularities are naked enough, it may be possible to perform the functional equivalent of a supertask in which one observer, by making use of the labor of another observer, can gain within a finite time a knowledge of the outcomes of an infinite number of operations. This would, for example, seemingly undercut the moral of Church's proof of

the undecidability of arithmetic since it would now be possible to obtain knowledge of the results of mechanically checking all potential proof sequences for a given formula.

Chapter 5 deals with a problem pertaining not so much to the big bang singularity itself as to the way in which the singularity is set in standard cosmological models. The so-called horizon problem turns out to be as much a problem about conditions for the explanatory adequacy of a scientific theory as it is a problem of physics. Claims of the advocates of inflationary cosmology to have solved the horizon problem are critically examined and, I hope, somewhat deflated.

Chapter 6 examines recent work in physics on time travel and time machines. It is argued that the philosophical literature on these topics, and especially that part of the literature on the "grandfather paradox" and similar paradoxes, has failed to diagnose what is genuinely problematic about closed timelike curves in relativistic spacetimes. An attempt is made to reorient the discussion in what I regard as a more fruitful direction.

Chapter 7 deals with a cluster of issues revolving around the notions of eternal recurrence and circular time, notions that have exercised a peculiar fascination on physicists and philosophers alike. It turns out to be more difficult than might have been guessed to pin down what these notions mean in terms of relativistic spacetime structure. Some of the ways in which eternal recurrence have been imagined to be possible are shown to be physically meaningless. Others are meaningful but are unrealizable according to GTR because of the occurrence of singularities. And the supposed dichotomy between open and closed times is shown to be not a dichotomy at all.

The closing chapter, chapter 8, does not attempt a "what does it all mean" summary; indeed, any such summary would be an insult to an enormously complex and fascinating subject. But it does address the antipathy to spacetime singularities, an antipathy that started with Einstein and is still found in some segments of the physics community.

Although I will from time to time have remarks about the implications of quantum gravity, the focus of this book is squarely on classical GTR. This is a severe limitation, but not an unnatural one if contentious speculation is to be avoided. And the problems of singularities and acausalities as they occur in classical GTR seem to me sufficiently interesting in their own right as to justify treating them on their own terms. But readers will have to judge for themselves.

Notes

1. An excellent survey of various approaches to the measurement problem is to be found in Albert (1992).

2. Or so it would seem; but see chapter 2.

3. A comprehensive and historically accurate account of Einstein's struggles with spacetime singularities, much less of singularities in general, has yet to be written.

Havas (1989, 1993) has dealt with Einstein's attitude as it related to the problem of motion. Eisenstaedt (1989, 1993) has provided an illuminating discussion of the evolving understanding of the Schwarzschild singularity. A good overview of work on singularities (up to 1980) is to be found in Tipler, Clarke, and Ellis (1980); but as the authors of this seminal article would surely agree, for historians of science the article is an invitation to do a careful history of this complex subject.

4. The final EFE can be written in two equivalent forms:

(i) $R_{ab} = 8\pi(T_{ab} - \frac{1}{2}Tg_{ab}) + \Lambda g_{ab}$

(ii) $R_{ab} - \frac{1}{2}Rg_{ab} + \Lambda g_{ab} = 8\pi T_{ab}$

Here g_{ab} is the spacetime metric, T_{ab} is the *energy–momentum tensor*, $T =: \mathrm{Trace}(T_{ab})$, R_{ab} is the *Ricci tensor* associated with g_{ab}, $R =: R_a^a$ is the *curvature scalar*, and Λ is the *cosmological constant*. From (i) one infers that when $\Lambda = 0$ the vacuum fields equations ($T_{ab} = 0$) take the form $R_{ab} = 0$. From (ii) one infers that they also take the form $G_{ab} =: R_{ab} - (1/2)Rg_{ab} = 0$, where G_{ab} is the *Einstein tensor*. By consistency, one infers that $R = 0$ for the vacuum case. Einstein (1915) started from the equations $R_{ab} = \text{const.} \times T_{ab}$.

5. Here and below I have taken the liberty of changing sign conventions to conform to a signature $(+++-)$ for the spacetime metric.

6. See Earman and Janssen (1993) for an account of Einstein's derivation of the perihelion shift for Mercury.

7. The translation is courtesy of John Norton.

8. I am indebted to Jean Eisenstaedt for pointing out to me how Hilbert's definition goes astray; see Eisenstaedt (1993). There are some subtleties here that can only be brought out with the help of the concept of an *extension* of one spacetime by another; see sections 2.3 and 2.5 of chapter 2.

9. The postcard is in the archives of the Sterrewach at Leiden. The English translation is from Kahn and Kahn (1975).

10. An account of this correspondence is given in Kerszberg (1989).

11. Here the sign of ds^2 is fiddled, if necessary, so that $ds \geq 0$.

12. John Stachel (1993) has noted that in addition to providing a non-singular version of the de Sitter metric, Lanczos also showed that sense could be made of the "mass-horizon" at $r = \pi R/2$ and in this way vindicated Einstein's claim that the de Sitter solution was not anti-Machian (see below).

13. As the reader will no doubt have gathered, in order to save complications in the notation, the same symbol r is being used in different line elements, although the physical interpretation is different in the different cases.

14. Finkelstein (1958) independently discovered the same coordinate transformation. The coordinates are thus referred to as the *Eddington–Finkelstein coordinates*.

15. English translation from Havas (1993). EA xx-yyy stands for Control Index item # xx-yyy in the Einstein Archive.

16. Einstein's attitude towards the cosmological constant is discussed below.

17. This point is emphasized by Torretti (1983, section 5.8).

18. I am indebted to John Norton for this point.

19. Here ∇_a stands for the covariant derivative operator associated with the metric g_{ab}. $\nabla_a G^{ab} = 0$ is a geometrical identity. Thus, $\nabla_a T^{ab} = 0$ is a consequence of the EFE.

20. For a modern derivation of the geodesic postulate see Geroch and Jang (1975).

21. Tipler et al. (1980) note that Einstein said that he had read the text of Peter Bergmann's book *Introduction to the Theory of Relativity* (1942) and that in this book there is no mention of the problem of singularities, save for a reference to Einstein's (1939) way of dealing with the Schwarzschild $r = 2M$ "singularity."

22. It was not until nearly four decades after Einstein's death that non-singular particle-like solutions were discovered for the non-vacuum EFE. It has been shown recently that the Einstein–Yang–Mills equations admit solutions that are static, asymptotically Minkowskian, singularity free, and possessing positive ADM mass; see Smoller, Wasserman, Yau, and McLeod (1991) and Smoller and Wasserman (1993). Such solutions would not have satisfied Einstein since he hoped to be able to represent particles without the help of a non-zero T^{ab} (see below).

23. Recently this problem has arisen again, and some cosmologists want to solve it by reintroducing Λ; see Crosswell (1993).

24. As John Stachel (private communication) has emphasized to me.

25. See Pais (1982) for an account of Einstein's work on unified field theories. Pais (1982) and Stachel (1986) contain remarks on Einstein's attempt to derive quantum constraints from classical field theory.

26. The same sentiment is expressed again a few pages later in the part of the paragraph immediately preceding the previous quotation; see Einstein (1955, p. 123).

27. Or is it too far fetched to suggest that Einstein might have reconsidered his opposition to QM if he had thought that the quantum might hold the key to getting rid of spacetime singularities?

28. Havas (1989) documents how Einstein ignored much of the work done by others on the problem of motion in the gravitational field. Einstein preferred to work things out for himself from what he regarded as first principles.

29. I am skipping over many of the details of the Einstein–Silberstein dispute. For the full story the reader will have to consult Havas (1993).

30. It is sometimes said that the singularity on the axis can be interpreted as "rods or struts" that hold the particles apart; see Kramer et al. (1980, section 17.1). As far as I know, this notion has never been made precise. The kind of singularity here is similar to the cone singularity discussed in section 2.2 of chapter 2.

31. John Stachel (private communication) has noted that prior to the final formulation of GTR Einstein raised the possibility of closed timelike curves in relativistic spacetimes but pronounced the possibility repugnant to his physical intuition: "Dies widerstrebt meinem physikalischen Gefühl aufs lebhaftteste" (Einstein 1914, p. 1079).

2

Defining, Characterizing, and Proving the Existence of Spacetime Singularities

2.1 Introduction

We saw in chapter 1 that shortly after GTR was put in its final form, attempts were made by Hilbert (1917) and Einstein (1918) to specify the conditions under which a spacetime is singular. Some small amount of progress was made on the problem over the next four decades. But as late as 1960 György Szekeres could accurately state that "an exact definition of what should be regarded as a true singularity of a [pseudo] Riemannian manifold has, to my knowledge, never been proposed" (Szekeres 1960, p. 285).[1] Over the following years rapid strides were made in proving the existence of singularities in solutions to EFE and in understanding what these singularities involved in terms of spacetime structure. But towards the end of the 1960s there were still major uncertainties in how to pin down the elusive notion of spacetime singularity. Some of these uncertainties were summarized in an important paper by Robert Geroch (1968a), part of which was cast as a Galilean dialogue, a form nicely chosen to reflect the unsettled state of the subject. In the intervening years our knowledge about these matters has increased tremendously, but uncertainties still remain.[2] The account which follows does not attempt to trace the historical development over recent decades but rather is aimed at providing an accessible introduction to these issues, issues which remain among the most exciting and difficult in the foundations of spacetime theories.

It will emerge that there are at least four distinct though interrelated concepts of spacetime singularities. The first can be seen as the natural heir of Einstein's (1918) idea that a singular spacetime is one in which the metric somehow breaks down at a finite distance. The second develops from an attempt to make precise the notion of "at a finite distance"; it promotes geodesic incompleteness or some generalization thereof as the essential characteristic of singularities. The third, which appeals to the notion of "missing points," is extensionally equivalent to the first two in many cases

but is nevertheless conceptually distinct. The fourth, which evolved from Roger Penrose's cosmic censorship hypothesis, departs sharply from the path set by Einstein's definition since it recognizes the presence of "naked singularities" even when the singularities fail to lie "at a finite distance." It will also emerge that there is a tension between the noun and adjective conceptions of spacetime singularities. The former attempts to conceive of singularities as entities that can be localized while the latter eschews localization and is content to speak of singular spacetimes when these spacetimes exhibit certain large-scale or global features.

2.2 What is a spacetime singularity?

In classical and special relativistic physics, the definition, if not the physical interpretation, of a singularity is straightforward. So for instance, a singularity in the electromagnetic field exists at a point p in Minkowski spacetime if the electromagnetic energy density "blows up" at p (i.e., increases beyond all bounds as one approaches p along any path). When we shift to the context of GTR two related differences make themselves felt. First, there is no longer a fixed spacetime background which by general agreement is non-singular and against which the blowup of a physical quantity can be measured. And secondly, we are now interested not in singularities in some physical field on the spacetime but rather in singularities in the spacetime *itself*. Nevertheless, we can try to carry over to GTR the idea that a singularity involves the blowup of a physical quantity.

First try

A singular point in the spacetime is a point where, say, the curvature of the spacetime blows up. Strictly speaking, this idea is incoherent since if the spacetime metric is ill-behaved at a point, then that point is not part of the spacetime. This attitude is codified in the usual definition of a general relativistic spacetime as a pair M, g_{ab} where M is a connected differentiable manifold (without boundary) and g_{ab} is a Lorentz metric which is defined and C^k $(k \geq 0)$ at every $p \in M$.[3]

Already at this early stage we have run into one of the perplexities that will eventually force some fundamental choices in the treatment of singularities. If we want to continue to speak of spacetime singularities in terms of singular points, then we have to recognize that the kind of existence these entities have (I will use ∃xistence to refer to it) is a ghostly one, for it does not conform to the wholly sensible slogan "To exist is to exist in space and time." Perhaps this ∃xistence can be made less ghostly by replacing it by regular existence in an augmented spacetime which contains singular points as ideal boundary points attached to the spacetime manifold. On the other hand, a failure to find a sensible procedure for the replacement would support

the position that noun constructions about singularities should be dropped in favor of adjectival constructions—talk of singularities as entities should be replaced by the less ontologically loaded talk of singular spacetimes.

Whichever way one comes down on this issue of singularities as "missing points," it seems sensible to try to develop a criterion for when a spacetime is to be counted as singular. One can try to accomplish this with a more sophisticated version of the blowup idea.

Second try

If the spacetime M, g_{ab} contains a suitable curve $\gamma \subset M$ along which the curvature becomes unboundedly large, then the singular nature of the spacetime is indicated. One difficulty here is that spacetime curvature is characterized by the Riemann tensor R_{abcd}. It will not do to take the singular nature of the spacetime to be indicated by the blowup of some component of the curvature tensor in some coordinate system, for that ill behavior may be the fault of the coordinate system and not of the Riemann tensor. We can avoid this problem by looking at curvature scalars such as $R\ (=R_a{}^a)$, $R_{ab}R^{ab}$, $R_{abcd}R^{abcd}$, and other scalar invariants formed by taking outer products of R_{abcd} and its derivatives and contracting indices with the help of g_{ab}.[4] The blowup of such scalars is surely a good indication of the singular nature of the spacetime (at least if suitable restrictions are placed on the curve γ along which the blowup occurs, as will be discussed below). The trouble is that although we have a sufficient criterion, we do not have a necessary one. Wald (1984a, p. 214) has noted that in some vacuum solutions to EFE describing gravitational plane waves rippling through spacetime, all such curvature scalars can vanish even though the curvature tensor itself is singular.[5]

Third try

It is too soon to give up on our basic idea. On our curve γ, choose an orthonormal tetrad field or *frame*[6] $e_i{}^a$, $i = 1, 2, 3, 4$, and use this frame to define the *physical components* $R_{(i)(j)(k)(l)} =: R_{abcd}e_i{}^a e_j{}^b e_k{}^c e_l{}^d$ of the curvature tensor. These physical components may be badly behaved if, for example, the frame is allowed to spin madly around. So let us require that the tetrad field be parallely propagated (p.p.) along γ.[7] If the physical components of the curvature tensor in such a p.p. frame blow up, we can say that the singular nature of the spacetime has been revealed.[8]

The more precise our trial criterion of singularity becomes, the more worries that arise. First, if the physical components of R_{abcd} in any p.p. frame along γ blow up but only as γ goes off to spatial or temporal infinity, then the singular nature of the spacetime has not been demonstrated in the sense intended by Einstein since singular behavior does not lie "at a finite distance" (see chapter 1). Sussmann (1988) has produced a family of spherically symmetric, shear free, perfect fluid solutions to EFE, some of whose members

contain a kind of "asymptotically delayed big bang": curvature scalars diverge along timelike and null geodesics that approach this delayed big bang, but a geodesic of infinite affine length is needed to reach the bang.[9] Clearly then some appropriate restriction on the length of γ is needed if curvature divergence along γ is to indicate a singularity in the sense intended by Einstein. Second, what if the physical components of R_{abcd} do not blow up but do fail to approach a limit? Should we recognize oscillatory as well as blowup behavior as an indicator of singularities? Third, the necessity of the trial criterion is also challenged by the following example of a cone singularity (Ellis and Schmidt 1977, p. 921; Wald 1984a, p. 214). In a polar cylindrical coordinate system, the Minkowski line element for $\mathbb{R}^4$ is

$$ds^2 = d\rho^2 + \rho^2\, d\phi^2 + dz^2 - dt^2, \qquad \rho =: (x^2 + y^2)^{1/2}, \qquad 0 \leq \phi \leq 2\pi.$$

Choose an angle $\phi_0 < 2\pi$, remove the wedge consisting of the points such that $0 < \phi < \phi_0$, and then identify the points $(\rho, 0, z, t)$ with the corresponding points (ρ, ϕ_0, z, t). All the points on the resulting manifold, save for those on $\rho = 0$, inherit a smooth metric from the Minkowski metric on $\mathbb{R}^4$. If we exclude the points $\rho = 0$ we arrive at a spacetime with a flat ($R_{abcd} = 0$) metric. But despite the fact that there is no ill behavior in the curvature, this spacetime can plausibly be classified as singular. Consider an observer who is born and then spends his life cruising along without acceleration. If his world line is aimed at (the missing) $\rho = 0$, then even if he drinks of the fountain of youth he will experience only a finite lifetime (as measured by proper time along his world line). Certainly the ghost of such an observer whose time has run out will feel justified in complaining that the spacetime is pathological.

Fourth try

Combining the first and third points above suggests that we set aside for the moment the idea that spacetime singularities involve curvature blowup and explore instead the alternative idea that spacetime singularities are associated with incomplete curves, curves that cannot be extended to an arbitrarily large value of some suitable parameter. Implementing the latter idea requires some care. A *half-curve* is a curve which has one endpoint and which is inextendible in the direction away from the endpoint.[10] A half-curve is said to be *complete* with respect to some parameter if the values of the parameter are unbounded. It will not do to count a spacetime as singular if it contains a timelike half-curve which is incomplete with respect to proper time, for if no restriction is placed on the acceleration of such a curve, then even Minkowski spacetime will be counted as singular.[11] The simplest course to take at this juncture is to focus on unaccelerated curves or geodesics. A geodesic half-curve is said to be *incomplete* just in case it has finite affine length. (For a timelike or spacelike geodesic this is equivalent to saying that its proper length is finite.)

It is easy to create artificial examples of spacetimes that are geodesically incomplete. For example, start with Minkowski spacetime $\mathbb{R}^4$, η_{ab}, remove a compact ball K, and restrict η_{ab} to $\mathbb{R}^4 - K$.[12] The resulting spacetime contains incomplete geodesics of all three kinds: timelike, spacelike, and null. In this case there is a straightforward sense in which the incompleteness is due to the existence of missing points. But by the same token, the singular nature of the surgically mutilated spacetime is non-intrinsic since it can be repaired by extending the spacetime back to the full Minkowski spacetime. Here it is appropriate to digress on the notion of extensions of spacetimes since the ideas involved are crucial to much of what follows.

2.3 Extensions of spacetimes

$\tilde{M}, \tilde{g}_{ab}$ is said to be an *extension* of M, g_{ab} just in case there is an isometric imbedding of the latter into the former, i.e., there is a diffeomorphism $\varphi: M \to \tilde{M}$ such that $\tilde{g}_{ab}|_{\varphi(M)} = \varphi^* g$.[13] The extension is *proper* just in case $\varphi(M)$ is a proper subset of $\tilde{M}$. M, g_{ab} is *properly extendible* just in case there is a proper extension of it. To illustrate some of the subtleties of these definitions, it will be useful to consider an example adapted from Wald (1984a, p. 149). Consider the submanifold $M \subset \mathbb{R}^2$ covered by the coordinates x, t such that $-\infty < x < +\infty$ and $0 < t < +\infty$. Define a metric g_{ab} on M whose line element is $ds_g^2 = dx^2 - dt^2/t^4$. This metric is extendible to all of $\mathbb{R}^2$ by means of the imbedding φ that takes a point with coordinates (x, t) to the point with coordinates $(x, 1/t)$. Introduce new coordinates x', t' such that $x' = x$ and $t' = 1/t$, i.e., the new coordinates of the new point are numerically the same as the old coordinates of the old point. So $(\varphi^* g)_{x'x'}(\varphi(x, t)) = 1$ and $(\varphi^* g)_{t't'}(\varphi(x, t)) = -1/t'^4$. Thus, $ds^2_{\varphi^* g} = dx'^2 - dt'^2/t'^4 = dx^2 - dt^2$. This is just the (two-dimensional) Minkowski metric, so of course there is a $\tilde{g}_{ab}$ defined on all of $\mathbb{R}^2$ such that $\tilde{g}_{ab}|_{\varphi(M)} = \varphi^* g$. But now consider a second imbedding of M into $\mathbb{R}^2$ which is just the natural inclusion map. Then relative to this imbedding there is no C^0 metric $\tilde{g}_{ab}$ defined on all of $\mathbb{R}^2$ such that $\tilde{g}_{ab}|_M = g_{ab}$ since $g_{tt} = -1/t^4$ blows up as $t \to 0^+$.

The example just given is relatively uninteresting from the point of view of singularities since the metric in question would not be considered singular (for the candidate singularity at $t = 0$ does not lie at a finite distance, e.g., all the geodesics that approach it are of infinite affine length). But an analogous moral holds for the Schwarzschild metric. Let $M \subset S^2 \times \mathbb{R}^2$ be the open submanifold covered by the Droste coordinates for $r > \alpha$ (see section 1.2). The points at $r = \alpha$ do lie at a finite distance (e.g., they can be reached by spacelike geodesics of finite length) and are candidate singularities since the g_{rr} component of the Schwarzschild metric blows up as r approaches α. And in fact under the imbedding that is given by the natural inclusion map for points labeled by the Droste coordinates, there is no C^0 extension of the metric to and beyond the Schwarzschild radius. (This is one way of expressing

what was right about Hilbert's claim as discussed in section 1.2) However, there is another imbedding due to Kruskal (1960) which does implement a C^∞ extension of the Schwarzschild metric to all of $S^2 \times \mathbf{R}^2$. (This expresses what is wrong about Hilbert's claim.) The details will be saved for section 2.5.

By contrast to the Schwarzschild radius ($r = \alpha$), there is no C^2 extension of the Schwarzschild metric to and through $r = 0$ since the Kretschmann curvature scalar $R_{abcd}R^{abcd}$ ($\sim 1/r^6$) blows up rapidly as $r = 0$ is approached. But do not be misled by this example. Constructing extensions can be prevented not only by ill behavior of the curvature but also by topological obstructions. Consider again the case of the cone singularity introduced at the end of section 2.2. The flat metric cannot be smoothly continued to $\rho = 0$. For in the plane z = constant, t = constant, a circle of radius ρ has circumference $(2\pi - \phi_0)\rho < 2\pi\rho$ for any $\rho > 0$, which is impossible if the origin $\rho = 0$ is part of a Euclidean space (Ellis and Schmidt 1977, p. 921).[14]

A spacetime is said to be *maximal* (with respect to some stipulated continuity/differentiability condition on the metric, e.g., C^2) just in case there is no proper extension (satisfying the stipulated condition). In all of the above examples of non-maximal spacetimes there is a maximal extension. Does this feature hold in general? The answer might seem to be obviously positive, e.g., every C^k spacetime can be extended to a maximal C^k spacetime. However, the "obvious" proof of this fact via a Zorn's lemma construction does not work.[15] What can be shown directly from Zorn's lemma is something weaker. An (n-dimensional) *framed spacetime* M, g_{ab}, F is a spacetime equipped with a distinguished orthonormal n-ad F of vectors at some point of M. Call $\tilde{M}, \tilde{g}_{ab}, \tilde{F}$ an extension of M, g_{ab}, F just in case there is an isometric imbedding φ of M into $\tilde{M}$ such that $\varphi^* F = \tilde{F}$. Then by a Zorn's lemma construction, every framed spacetime can be extended to a (not necessarily unique) maximal framed spacetime (Geroch 1970a). The desired result now follows easily. Let M, g_{ab} be an arbitrary C^k spacetime. Frame it in some way to form M, g_{ab}, F. Obtain a maximal framed extension $\tilde{M}, \tilde{g}_{ab}, \tilde{F}$. Then $\tilde{M}, \tilde{g}_{ab}$ is a (not necessarily unique) maximal extension of M, g_{ab}.

Metaphysical considerations suggest that to be a serious candidate for describing actuality, a spacetime should be maximal. For example, for the Creative Force to actualize a proper subpart of a larger spacetime would seem to be a violation of Leibniz's principles of sufficient reason and plenitude. If one adopts the image of spacetime as being generated or built up as time passes then the dynamical version of the principle of sufficient reason would ask why the Creative Force would stop building if it is possible to continue. However, this image does not sit well with the four-dimensional way of thinking, and in any case it runs into trouble in its own terms: since extensions of spacetime are generally non-unique there may be many ways to continue building and the Creative Force may be stymied by a Buridan's ass choice (see Clarke 1993, pp. 8–9). Some readers may be shocked by the introduction of metaphysical considerations in the hardest of the "hard sciences." But in fact leading workers in relativistic gravitation, though they don't invoke the

name of Leibniz, are motivated by such principles (see, for example, Geroch 1970a, p. 262; Penrose 1969, p. 253). The intrusion of metaphysics will also be evidenced in all of the chapters to follow.

2.4 The received definition of singularities

Any non-maximal spacetime is geodesically incomplete. But such a spacetime is arguably not an acceptable model for actual spacetime. So suppose that we focus on maximal spacetimes. If there are any singularities in such a spacetime they are essential or irremovable. The task now is to find a criterion that will signal when a spacetime is singular. Will geodesic incompleteness suffice? An appeal to the analogy of the better understood case of a Riemannian space would seem to encourage a positive answer.

A *Riemannian space* M, h_{ab} (where h_{ab} is a positive definite Riemannian metric as opposed to a pseudo-Riemannian metric of relativistic spacetimes)[16] can be made into a metric space by a standard construction. Choose an arbitrary pair of points $p, q \in M$ and consider all piecewise differentiable curves $\gamma(v)$ joining p and q, where the curve parameter v is arbitrary. The length of $\gamma(v)$ between p and q is $\int_p^q (h(d\gamma/dv, d\gamma/dv))^{1/2}\, dv$. Define the distance $d(p, q)$ between p and q to be the greatest lower bound on the length of such curves. The following properties of this distance function can be proved: for any $p, q, r \in M$

(i) $d(p, q) = d(q, p)$

(ii) $d(p, q) = 0$ if and only if $p = q$

(iii) $d(p, q) \leq d(p, r) + d(r, q)$

The existence of a distance function with these three properties is what is meant by saying that $M, d(\ ,\ )$ is a *metric space*. Further, it can be shown that the manifold topology is compatible with the metric topology in that

(iv) the metric balls $B(p, \varepsilon) =: \{q \in M: d(p, q) < \varepsilon\}$ for all $p \in M$ and all $\varepsilon > 0$ form a basis for the topology of M.

For a metric space we can define the notion of a *Cauchy sequence* of points p_i, $i = 1, 2, 3, \ldots$, by the condition that for any $\varepsilon > 0$ there is an $\mathcal{N}$ such that for any $m, n > \mathcal{N}$, $d(p_m, p_n) < \varepsilon$. A metric space is said to be *Cauchy complete* just in case every Cauchy sequence converges to some point in the space. For metric spaces then, Cauchy incompleteness is the signal of the ∃xistence of missing points. Furthermore, this peculiar sense of ∃xistence can be replaced by ordinary existence in a larger space. For it is a theorem that any Cauchy incomplete metric space can be isometrically imbedded as an open dense subset of a complete metric space. Roughly, the new points of the larger space (= original space plus the missing points of the incomplete space) are constructed from equivalence classes of non-convergent Cauchy sequences,

where two Cauchy sequences p_i and q_i are taken to be equivalent just in case $d(p_i, q_i) \to 0$ as $i \to \infty$. (In just this way the real numbers can be constructed by taking the Cauchy completion of the rationals.)

Returning now to the case of a relativistic spacetime M, g_{ab} with a pseudo-Riemannian metric, we are unable to avail ourselves of the metric space characterizations of completeness and incompleteness. For in M, g_{ab} any pair of points $p, q \in M$ can be joined by a broken null curve, with the result that with $d(p, q)$ defined as before, the distance between the points is always zero. However, in the Riemannian case the Hopf–Rinow theorem shows that Cauchy completeness is equivalent to geodesic completeness. Thus, if the analogy between Riemann spaces and relativistic spacetimes holds, it would seem reasonable to use geodesic completeness as a criterion of completeness in relativistic spacetimes.

The first problem we face in implementing the criterion is that in relativistic spacetimes there are three different kinds of geodesics—timelike, spacelike, and null—and it has been shown by explicit examples that the various kinds of geodesic incompleteness are inequivalent. Indeed, there is complete inequivalence in that one can have each form of incompleteness in the presence of the other two forms of completeness, as well as any two forms of incompleteness in the presence of the third form of completeness. To give a feeling for how the various forms of incompleteness can come apart it will be useful to sketch Geroch's (1968a) example of a spacetime that is timelike geodesically incomplete while being spacelike and null geodesically complete. Start with two-dimensional Minkowski spacetime $\mathbb{R}^2$, η_{ab} and generate the conformally related spacetime $\mathbb{R}^2$, $\Omega^2\eta_{ab}$ where $\Omega^2\colon \mathbb{R}^2 \to (0, +\infty)$ is a C^∞ map with the following properties. For some fixed inertial coordinate system (x, t):

(1) $\Omega^2(x, t) = 1$ if $x \leq -1$ or $x \geq 1$

(2) $\Omega^2(x, t) = \Omega^2(-x, t)$ for all x, t

(3) $\Omega^2(0, t)$ goes rapidly to 0 as $t \to \infty$

The third property implies that the timelike half-geodesic that starts at $(0, 0)$ and then traces the t axis forward or backward in time is incomplete. On the other hand, any inextendible null or spacelike geodesic that starts in or enters the strip $-1 \leq x \leq 1$ must eventually escape, and thus by property (1) is complete.

Because we can imagine ourselves travelling along a timelike geodesic, timelike geodesic incompleteness is the most psychologically disturbing of the three forms of incompleteness. Nevertheless there are spacetimes which we would want to count as singular but which are timelike geodesically complete. Reissner–Nordström spacetime, which describes the exterior field of a spherically symmetric, electrically charged body, is one example (see Hawking and Ellis 1973, pp. 156–161). Intuitively, the singularity, which can be reached by a null or a spacelike geodesic, exerts a repulsive force that causes free-falling

massive particles to miss hitting it. Though less psychologically disturbing than either timelike or null incompleteness, spacelike geodesic incompleteness cannot be ignored either. So to be on the safe side, let us try on for size this *tentative definition*: a spacetime is *non-singular* just in case it is geodesically complete in all three senses.

An example due to Geroch (1968a) brings into question the sufficiency of this definition. Consider a timelike half curve (not necessarily geodesic) $\gamma(\tau)$ parameterized by its proper time τ. At each point on the curve construct the normalized tangent vector $V^a(\tau)$ $(g_{ab}V^a(\tau)V^b(\tau) = -1)$. The four-acceleration is then defined as $A^a(\tau) =: V^b(\tau)\nabla_b V^a(\tau)$. If $\gamma(\tau)$ is a geodesic, then $A^a(\tau) = 0$ for every τ. But instead of considering only timelike geodesics, let us consider the more general class of curves of bounded acceleration, i.e., there is a positive number B such that the magnitude of acceleration $a(\tau) \leq B$ for all τ, where $a(\tau) =: (A^b(\tau)A_b(\tau))^{1/2}$. If in addition such a curve has finite proper length, then one could in principle construct a rocket ship whose world line would instantiate this curve using only a finite amount of fuel, and the cosmonaut at the controls would biologically age only a finite number of years. But by hypothesis, the curve cannot be extended, so the cosmonaut does not enjoy the possibility of prolonging his life. The unfortunate fellow would surely feel justified in regarding the spacetime he inhabits as being singular. Now the point of Geroch's example is that such behavior can occur even if the spacetime is geodesically complete. It seems prudent then to strengthen the proposed definition of a non-singular spacetime to require not only geodesic completeness but also *bounded acceleration completeness*, that is, that every timelike half-curve of bounded acceleration has infinite proper length.

But even this modification of our tentative definition may not suffice, for there is a still more general sense of completeness that must be taken into account. Consider a half-curve $\gamma\colon [0, v_+) \to M$, $v_+ \leq +\infty$. Choose an orthonormal basis $e_i^a(0)$ for the tangent space at $\gamma(0)$ and parallel transport this basis along $\gamma(v)$ to give the frame field $e_i^a(v)$, $v \in [0, v_+)$. The tangent vector $\mathbf{V} = (\partial/\partial v)_{\gamma(v)}$ to the curve can then be written as $V^a = \sum_{i=1}^4 X^i(v)e_i^a(v)$. And the *generalized affine parameter* (g.a.p.) of $\gamma(v)$ is defined by $\lambda =: \int_0 (\sum_{i=1}^4 (X^i(v))^2)^{1/2}\, dv$. The choice of a different frame field $e'^a_i(v)$ will lead to a different g.a.p. λ' on $\gamma(v)$; but it can be shown that $\gamma(v)$ will have finite λ' length if and only if it has finite λ length. Thus, the notion of finite (or infinite) *generalized affine length* (g.a.l.) for a half-curve is a well-defined notion.

It is easily seen that if γ is a geodesic, then the g.a.p.s on γ will be affine parameters; but, of course, the g.a.p. is defined for all C^1 curves, not just for geodesics. For an arbitrary timelike curve (geodesic or not), if the curve has infinite proper length, then it has infinite g.a.l. But a timelike curve of finite proper length may not have finite g.a.l. if, for example, the curve experiences unbounded acceleration. A sufficient condition for the g.a.l. incompleteness of a timelike curve has been developed by Sussmann (1988). Consider a

timelike half curve $\gamma(\tau)$, $\tau \in [0, \tau_1)$, $\tau_1 < +\infty$, where τ is proper time along the curve. Assuming that this curve has finite proper length, it will also have finite g.a.l. if there exist an L and a τ_2 such that $0 < L < 1$, $0 \leq \tau_2 \leq \tau_1$, and $a(\tau)|\tau_1 - \tau| < L$ for all $\tau_2 \leq \tau \leq \tau_1$. Clearly, if $a(\tau)$ is bounded along $\gamma(\tau)$, this criterion is met.

A spacetime is said to be *b-complete* just in case every half-curve has infinite g.a.l.[17] From the above remarks it follows that *b*-completeness entails geodesic completeness and bounded acceleration completeness. But *b*-completeness is stronger in that it is not entailed by the other senses of completeness. We can now formulate the most widely accepted definition of a singular spacetime, the *semiofficial definition*: a spacetime is *singular* if and only if it is *b*-incomplete. (Hawking and Ellis (1973) take this criterion to apply to all curves. However, one might reasonably want to restrict it to timelike and null curves.) This semiofficial definition passes one quick reality check: it counts Minkowski spacetime as non-singular. Also according to the semiofficial view, the asymptotically delayed big bang in Sussmann's (1988) solutions (see section 2.2 above) is not counted as a singularity since any curve which reaches it is *b*-complete.

Each *b*-incomplete curve in a spacetime M, g_{ab} can be thought of as defining an ideal point on a boundary $\partial_b M$ of M. To give some life to the word 'boundary' here, we need a prescription that will tell us when two *b*-incomplete curves define the same boundary point; and, more ambitiously, we would like a prescription that will give some structure to the enlarged manifold $\hat{M} =: M \cup \partial_b M$—at a minimum, a topological structure, so that we can say which points in the boundary $\partial_b M$ are in the neighborhoods of what points in the interior M. We would then have succeeded in replacing the ∃xistence of spacetime singularities with existence in $\partial_b M$ in a manner that localizes the singularities. Schmidt's (1971) prescriptions use the fact that the bundle of orthonormal frames $O(M)$ admits a natural positive definite metric.[18] Thus, we know from the above remarks that $O(M)$ can be made into a metric space. It turns out that this metric space is Cauchy complete just in case the spacetime is *b*-complete. If M, g_{ab} is *b*-incomplete, the Cauchy completion can be taken in $O(M)$ and then projected down to give $M \cup \partial_b M$. Unfortunately, the resulting bundle boundary has some counterintuitive properties. For instance, in the FRW spacetimes $\partial_b M$ consists of a single point that is not Hausdorff separated from points of M (see Johnson 1977).

Such results might be taken to call into doubt the wisdom of defining singularities in terms of *b*-completeness. For if one goes beyond geodesic completeness and bounded acceleration completeness in order to guarantee freedom from singularities, why settle on *b*-completeness rather than on some other stronger or weaker notion of completeness? Delivering an appealing way to attach singular boundary points to the spacetime manifold would have provided an answer, but this is just what the *b*-boundary approach fails to deliver. However, other results tend to absolve the *b*-boundary approach for the failure. For instance, Geroch, Liang, and Wald (1982) have argued that

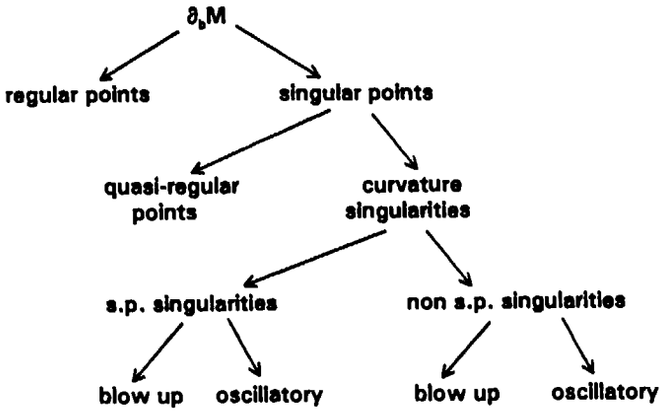

Fig. 2.1 The Ellis–Schmidt classification of spacetime singularities in the *b*-boundary approach

the failure of singular points to be Hausdorff separated from interior points is an inescapable feature of any attempt which seeks to represent singular points of a spacetime M, g_{ab} in terms of a topological space $\hat{M}$ which has M as an open dense subset and which shares with the *b*-boundary approach certain natural conditions; namely, any incomplete geodesic in M, g_{ab} terminates in a point of $\hat{M} - M$, and (in a certain technical sense) $\hat{M}$ is geodesically continuous.[19] It appears then that the goal of localizing spacetime singularities is unattainable without paying the price of counterintuitive features of the localization.

Another consideration in favor of the *b*-boundary approach is the interesting classification of singularities to which it gives rise. Following Ellis and Schmidt (1977), spacetime singularities can be put into the categories listed in Fig. 2.1. A *regular point* $p \in \partial_b M$ in the Ellis–Schmidt classification is what was called above an inessential singularity, i.e., $\tilde{M}, \tilde{g}_{ab}$ can be extended to a spacetime $\tilde{M}, \tilde{g}_{ab}$ such that the image of p is a point of $\tilde{M}$. *Singular points* are then defined as all of the non-regular points of $\partial_b M$. These points are then further subdivided into quasi-regular points and curvature singularities.

Let γ be a *b*-incomplete curve terminating in a point $p \in \partial_b M$, i.e., there is a generalized affine parameter v for γ and a $v_+ < +\infty$ such that $\gamma(v) \subset M$ for $v \in [0, v_+)$ and $\gamma(v_+) = p$. Suppose that in any p.p. frame on γ some of the physical components $R_{(i)(j)(k)(l)}(\gamma(v))$ do not approach limits as $v \to v_+$. Then p is said to be a *curvature singularity*. On the other hand, suppose that for any *b*-incomplete $\gamma(v)$, $v \in [0, v_+)$ terminating in the point $p \in \partial_b M$ there is some p.p. frame along $\gamma(v)$ such that all of the physical components of the curvature tensor approach limits as $v \to v_+$. Then p is said to be a *quasi-regular point*. This terminology is justified by a pretty result of Clarke (1973). Say that the *b*-incomplete γ is *locally extendible* just in case there is an open $U \subset M$ containing $\gamma([0, v_+))$ and an isometric imbedding φ of U into a $\tilde{U}, \tilde{g}_{ab}$

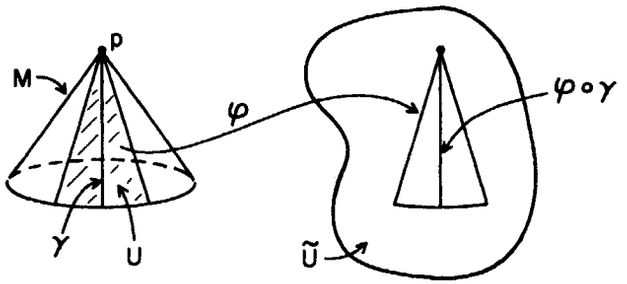

Fig. 2.2 An example of a spacetime that is locally extendible but not extendible (after Ellis and Schmidt 1977)

such that $\varphi \circ \gamma$ is C^0 extendible in $\tilde{U}$ beyond $v = v_+$. Clarke showed that if γ is a b-incomplete curve that corresponds to a quasi-regular singularity, then γ is locally extendible.

An example of a spacetime that is not extendible but is locally extendible is the spacetime with the conical singularity discussed in section 2.2. The idea is illustrated in Fig. 2.2. The image of the curve γ can be extended into $\tilde{U} - \varphi(U)$. U can be enlarged to cover "almost all" of the cone—indeed, all of the cone save a line extending down from the vertex. The boundary of the image $\varphi(U)$ of this U in $\tilde{U}$ contains two lines which meet at the vertex, whereas the pre-images of these lines on the cone consist of a single line. Thus, adding the vertex p back to form a global extension of the cone is prevented by there not being enough directions at p (Ellis and Schmidt 1977, p. 931). This example is admittedly artificial. But similar singularities occur naturally, as in the Curzon bipolar solution to EFE (recall section 1.3). Vickers (1987) has argued that two-dimensional cone singularities of this type model cosmic strings.

Another and much more perplexing example of a quasi-regular singularity is provided by Taub–NUT spacetime. Some of the causal features of this spacetime are captured in the simpler two-dimensional spacetime of Misner (1967); see Fig. 2.3. There is a null geodesic γ that starts on Σ and that winds round and round the Taub portion of the universe, approaching but never crossing the null surface $\mathcal{N}$. Since in the future direction γ uses up only a finite amount of affine parameter it defines a point in $\partial_b M$. However, a computation shows that the curvature components in a p.p. frame on γ remain well behaved as $\mathcal{N}$ is approached; so by Clarke's result there is a local extension in which γ can be continued.

Curvature singularities can be further classified, the first main subcategory being *scalar polynomial (s.p.) singularities*. There are in turn two main subcategories. (i) *Blowup s.p. singularities*. This is the case that is brought most immediately to mind by the phrase "spacetime singularity." Such a singularity occurs when there is a b-incomplete curve $\gamma(v)$, $v \in [0, v_+)$, defining a $p \in \partial_b M$ and a scalar curvature polynomial χ such that $\chi(\gamma(v))$ is not

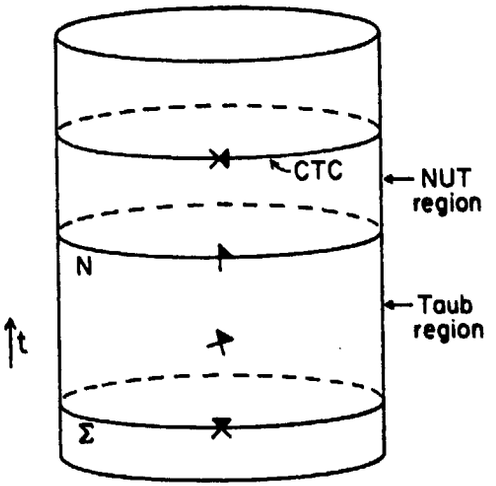

Fig. 2.3 Misner's two-dimensional spacetime illustrating some of the features of Taub–NUT spacetime

bounded as $v \to v_+$. Siklos (1979) proved that an s.p. blowup occurs just in case some of the physical components of the Riemann tensor are unbounded in *all* frames along γ, not just the p.p. ones. The FRW big bang models and the Kruskal–Schwarzschild model are examples where there are such s.p. blow up singularities. (ii) *Oscillatory s.p. singularities*. Here every s.p. remains bounded along $\gamma(v)$ but some s.p. $\chi(\gamma(v))$ fails to approach a limit as $v \to v_+$. A modified version of the Taub–NUT spacetime can be used to give an example of this type of behavior. Consider an incomplete null geodesic $\gamma(v)$ (with v an affine parameter) that winds its way up through the Taub region and approaches the null surface $\mathcal{N}$ that separates the Taub region from the NUT region. Since the curvature components in a p.p. frame along $\gamma(v)$ approach finite limits as $v \to v_+$, so a fortiori must any s.p. But one also has an independent argument for the latter fact. Since the metric is smooth in a compact neighborhood of $\mathcal{N}$, any s.p. must remain bounded in such a neighborhood. Furthermore, every point $p \in \mathcal{N}$ is a limit point of γ. Choosing on γ an infinite subsequence of points that converges to p, it follows that the values of any s.p. at these points must approach a finite limit as p is approached. And since $\mathcal{N}$ is a surface of homogeneity, the limit is the same at every $p \in \mathcal{N}$. Thus, any s.p. $\chi(\gamma(v))$ must approach a limit as $v \to v_+$. Now modify the example by choosing a $q \in \mathcal{N}$ and multiplying the metric by a conformal factor Ω^2 in a small neighborhood of q. The causal structure of (a two-dimensional version of) the resulting spacetime is still as pictured in Fig. 2.3, and relative to the new affine parameter $\tilde{v}$ of the conformal metric, $\gamma(\tilde{v})$ is still future incomplete. Furthermore, by the same argument as before, all of the scalar curvature polynomials, calculated in the new conformal metric,

must remain bounded as γ approaches $\mathcal{N}$. But the homogeneity of $\mathcal{N}$ has been destroyed by the conformal factor Ω^2, and with appropriate behavior of Ω^2 some s.p. can be made to oscillate without limit as $\tilde{v} \to \tilde{v}_+$.[20] Of course, this is merely a mathematical example since the conformally modified Taub–NUT spacetime is no longer a vacuum solution to EFE or even to any reasonable field equation. In a more physical vein, King (1974) has found oscillating s.p. singularities in cylindrically symmetric stationary dust solutions to EFE.

The second main category of curvature singularities consists of the *non-scalar polynomial singularities*. As with the first category, these can be divided into blowup and oscillatory types. In the blowup case the trouble can be attributed to something going wrong with parallel transport. For by definition, some physical component of the curvature tensor is unbounded in any p.p. frame along a b-incomplete curve γ while (by Siklos' result) all components remain bounded in some non-p.p. frame. Although various examples of non-s.p. (aka *whimper*) singularities can be given (see Siklos 1981) there are results that indicate that these singularities can occur only in special and, perhaps, physically unrealistic circumstances (see Ellis and Schmidt 1977).

Still further refinements of the classification of curvature singularities can be contemplated. For example, the ill behavior of the curvature can be traced to either the Ricci tensor or the Weyl tensor.[21] Or the curvature singularities can be ranked by their strength—for instance, as to whether they imply that a volume element is squeezed to naught as the singularity is approached (see chapter 3).

Although brief and incomplete, this glimpse of the varieties of spacetime singularities should suffice to indicate that only in special cases can singularities (qua incompleteness) in general relativistic spacetimes be made to conform to our original crude idea of some quantity blowing up.

2.5 The missing missing points

The analysis of singularities in terms of b-incompleteness leads to a pleasing classification scheme. But there is a disturbing feature of the analysis; namely, it is not true to an idea that is arguably a touchstone of singularities in relativistic spacetimes: spacetime singularities correspond to missing points. To see where contact is lost with this touchstone, return to the better understood case of a Riemannian space M, h_{ab}. We saw that h_{ab} generates a distance function $d(\ ,\ )$ that makes $M, d(\ ,\ )$ a metric space, and that Cauchy completeness for $M, d(\ ,\ )$ is equivalent to geodesic completeness for M, h_{ab}. In turn each of these completeness properties is equivalent to the property of *finite compactness*: every subset of M that is d-bounded (i.e., $\sup\{d(p, q): p, q \in X\} < \infty$) has compact closure. It follows immediately that any compact Riemann space is complete. This is essential to the idea of using

incompleteness as a criterion of missing points since a compact M is never a proper submanifold of another (connected, Hausdorff) manifold.

In the relativistic spacetime setting the trouble is that a spacetime M, g_{ab} can be b-incomplete (indeed, geodesically incomplete) even though M is compact. Misner (1963) provided an example of a two-torus T^2 equipped with a Lorentz metric that is null geodesically incomplete. There is also the example of Taub–NUT spacetime studied in the preceding section. Here the spacetime is not compact, but there are incomplete null and timelike geodesics that are contained in a compact set. In the two-dimensional Misner spacetime pictured in Fig. 2.3, the incomplete null geodesic γ is confined to the closure of a neighborhood of the surface $\mathcal{N}$. As a result, the spacetime manifold M cannot be extended to a larger $\tilde{M}$ in which the incomplete γ is continued (see Hawking and Ellis 1973, p. 289). In this example, however, the incomplete geodesic γ defines a quasi-regular point and is, therefore, locally extendible. But we also saw that a conformal modification of the metric turns this into an example in which γ is not even locally extendible.

Several reactions to these examples are possible. The first is to hold fast to the idea of missing points as the touchstone of spacetime singularities and, consequently, to modify the definition of singularities along the following lines.[22] *Proposed modification of the semiofficial definition*: M, g_{ab} is *non-singular* just in case every half-curve is b-complete or else is contained in a compact subset of M. One possible objection to such a definition is that it contains an element of arbitrariness in that one could construct related but different definitions (see Geroch 1968a). For instance, there are presumably cases where a b-incomplete curve is not contained in a compact set but does continually reenter a compact set. Should we regard such a spacetime as being singular or not? This particular challenge can be met by the answer "Non-singular, in order to be true to the 'missing points' touchstone." For the fact that the curve continually reenters a compact set $K \subset M$ means that it has a limit point $p \in M$, and this is enough to show that M cannot be extended to a larger $\tilde{M}$ into which the curve can be continued. But there may be other challenges along these lines that do not have such a clear answer.

A second reaction would be to insist that b-incompleteness, whether or not it makes contact with the missing points touchstone, does correspond to one good sense of spacetime singularity. Thus, speaking partly to examples of the sort discussed immediately above, Hawking and Ellis acknowledge "There is no possibility of the incompleteness having arisen from the cutting out of singular points." But they add:

> Nevertheless, it would be unpleasant to be moving on one of the incomplete timelike geodesics for although one's world line never comes to an end and would continue to wind round and round inside the compact set, one would never get beyond a certain time in one's life. It would, therefore, seem reasonable to say that such a spacetime is singular. (Hawking and Ellis 1973, p. 261)

For those who share this attitude some subscripting of the word 'singularity' is needed: singular$_1$ to refer to singularities associated with incompleteness, singular$_2$ to refer to singularities associated with missing points.

A middle way is also possible. It acknowledges that ideas of missing points and incompleteness lead to different concepts of spacetime singularities, but it claims that these concepts are nearly extensionally equivalent for physically reasonable cases. For apparently the only cases where the concepts can come apart are (1) for a b-incomplete spacetime M, g_{ab} with compact M, or else (2) for a spacetime M, g_{ab} where M is non-compact but a b-incomplete curve is contained in a compact $K \subset M$ or else continually reenters K. In the first case the spacetime contains a closed timelike curve. This follows from the fact that the sets $I^+(p)$, $p \in M$, $I^+(p) =: \{q \in M: p \ll q\}$, where $p \ll q$ means that there is a non-trivial future directed timelike curve from p to q, form an open cover of M. Since M is by assumption compact, there is a finite subcover $I^+(p_1), I^+(p_2), \ldots, I^+(p_n)$. Then either $p_1 \in I^+(p_i)$ for some $i > 1$ or not. If so, then $I^+(p_1)$ can be omitted. If not, $p_1 \in I^+(p_1)$ and there are closed timelike curves. Continuing in this way produces the desired conclusion. In the second case, if the incomplete curve is non-spacelike, the spacetime contains almost closed causal curves. More precisely, there is a violation of the *strong causality condition* which holds for a spacetime M, g_{ab} just in case for all $p \in M$, any open neighborhood $\mathcal{N}(p)$ of p contains a subneighborhood $\mathcal{N}'(p)$ which no non-spacelike curve reenters after leaving. For a proof that the strong causality condition is violated if a future inextendible causal curve is totally or partially imprisoned in a compact set, see Hawking and Ellis (1973, Prop. 6.4.7). There is a loophole here concerning a b-incomplete spacelike curve that is partially or totally imprisoned in a compact set, but assuming that this loophole can be plugged, the ground is laid for the following stance. Closed or almost closed causal curves (so the argument goes) make a spacetime a priori unacceptable as an arena for physics. And within the acceptable arenas, the two concepts of spacetime singularity—based respectively on the ideas of missing points and incompleteness—are in agreement. The a priori postulation of causal features was once a popular move. Of late, however, its popularity has slipped badly among general relativists. This matter will be discussed in section 2.6 and in more detail in chapter 6.

Even if in physically reasonable cases there is an extensional equivalence of the concepts of singularities as involving incompleteness and singularities as involving missing points, the two concepts are distinct, and the latter deserves to be developed in its own right. The modification proposed above for the semiofficial definition of singularities is ad hoc and does not do justice to the missing points idea. More justice can be done by employing some of the ideas of Susan Scott (1992), which I proceed to sketch below.[23]

If there are missing points for a spacetime M, g_{ab} they have to arise from deleting points from an envelopment of M. Formally, an *envelopment* of the (n-dimensional) M consists of a connected (n-dimensional) $\tilde{M}$ and an imbedding $\varphi: M \to \tilde{M}$. $\varphi(M)$ is an open subset of $\tilde{M}$. Taking the

closure of $\varphi(M)$ in M and subtracting off $\varphi(M)$ gives the *boundary* $B(M, \tilde{M}, \varphi) =: \overline{\varphi(M)} - \phi(M)$ of M in the envelopment $\tilde{M}, \varphi$. If $p \in B(M, \tilde{M}, \varphi)$ is not the limit point[24] of some appropriately incomplete half-curve of M, g_{ab} (e.g., an incomplete half-geodesic or a half-curve of finite g.a.l.) then p is said to be a *point at infinity*; otherwise p is said to be *at a finite distance*.

Focusing now on boundary points at a finite distance, we need to weed out ones that do not correspond to genuine singularities. The weeding is done in two steps. First, we have to pull out the regular points. Assume that the metric g_{ab} on M is C^k ($k \geq 1$). A point $p \in B(M, \tilde{M}, \varphi)$ is said to be C^l *regular* ($k \geq l \geq 1$) just in case there is an open submanifold $\hat{M} \subset \tilde{M}$ with $\varphi(M) \subset \hat{M}$ and $p \in \hat{M}$, and a C^l metric $\hat{g}_{ab}$ on $\hat{M}$ such that $\hat{M}, \hat{g}_{ab}$ is an extension of M, g_{ab} under φ. The second bit of weeding focuses on the non-regular boundary points that lie at a finite distance—the candidate singularities—and is more complicated because it takes into account the various ways the base manifold M can be enveloped. Let $\tilde{M}', \tilde{\varphi}'$ be a second envelopment, and let $B(M, \tilde{M}', \varphi')$ denote the boundary set for this second envelopment. $B' \subset B(M, \tilde{M}', \varphi')$ is said to *cover* $p \in B(M, \tilde{M}, \varphi)$ just in case for every open neighborhood $U' \subset \tilde{M}'$ of B', there is an open neighborhood $U \subset \tilde{M}$ of p such that $\varphi' \circ \varphi^{-1}(U \cap \varphi(M)) \subseteq U'$. The candidate singular point $p \in B(M, \tilde{M}, \varphi)$ is said to be *removable* just in case there is another envelopment $\tilde{M}'$, $\tilde{\varphi}'$ and a C^l regular subset $B' \subset B(M, \tilde{M}', \varphi')$ that covers p. A relevant example has already been provided by the discussion of the Schwarzschild metric in section 2.3. Take $M \subset S^2 \times \mathbb{R}^2$ to be the open subset covered by the Droste coordinates for $r > \alpha$ (recall section 1.2.). Under the imbedding of M into $\tilde{M} = S^2 \times \mathbb{R}^2$ given by the identity map provided by the Droste coordinates, the points $(r = \alpha, \theta, \phi, t)$ of $\tilde{M}$ belong to $B(M, \tilde{M}, \mathrm{id})$. These points are not C^l regular for $l \geq 0$, but they lie at a finite distance since they are the endpoints of incomplete geodesics in M. Thus, they are candidate singular points. However, they are removable. A second envelopment is given with the help of the Kruskal coordinates

$$X = (1/2)\sqrt{r - \alpha}\exp(r/\alpha)[\exp(t/2\alpha) + \exp(-t/2\alpha)],$$

$$T = (1/2)\sqrt{r - \alpha}\exp(r/\alpha)[\exp(t/2\alpha) - \exp(-t/2\alpha)].$$

The new imbedding φ' is given by sending (r, θ, ϕ, t) to (X, θ, ϕ, T). The boundary points $B(M, \tilde{M}', \varphi')$, where $\tilde{M}' = \tilde{M} = S^2 \times \mathbb{R}^2$, are C^∞ regular and cover the boundary points $(r = \alpha, \theta, \phi, t)$ of $B(M, \tilde{M}, \mathrm{id})$ (see Scott 1992, p. 184). By contrast, $r = 0$ corresponds to a non-removable singularity.

The present approach counts the Misner compact spacetime and the Taub–NUT spacetime as non-singular. It therefore reproduces in a motivated way the results of the ad hoc modification of the semiofficial definition considered above. Whether or not this approach will prove to be fruitful depends upon how it can be used to classify the various types of singularities

it considers to be genuine (i.e., non-removable). This matter cannot be pursued here (see Scott 1992 for some details).

2.6 Naked singularities

Penrose's (1969) cosmic censorship conjecture supposes that GTR has built in prohibitions against "naked singularities." This conjecture will be discussed in detail in chapter 3, but in advance it is worth pointing out that cosmic censorship raises a set of concerns that, while not orthogonal to those of the preceding sections, are at least oblique to them. The strictest form of cosmic censorship rules out naked singularities by requiring that the spacetime be *globally hyperbolic*, which is the technical way of saying that the spacetime does not contain any pathologies that prevents the implementation of global Laplacian determinism. A globally hyperbolic spacetime contains a *Cauchy surface*, i.e., a spacelike hypersurface Σ that intersects every causal curve without endpoint exactly once. The specification of appropriate initial data on Σ determines throughout the spacetime a unique solution to the coupled Einstein–matter equations, as will be discussed in chapter 3.

What needs to be emphasized here is that a spacetime can be globally hyperbolic and still singular in one or more of the senses studied above—the FRW big bang models are prime examples. And in the other direction the spacetime can fail to be globally hyperbolic (i.e., can be nakedly singular in Penrose's sense) without being singular in any of the senses studied above. To see this, note that global hyperbolicity is equivalent to the conjunction of two conditions. The first is strong causality, which we met in the preceding section. The reader can easily construct examples where strong causality fails but there are no incomplete curves or missing points. The second condition requires that $J^-(p) \cap J^+(q)$ be compact for $p, q \in M$, where $J^-(p) =: \{s \in M: s < p\}$ and $J^+(q) =: \{s \in M: q < s\}$ and $x < y$ means that there is a future directed non-spacelike curve from x to y. Construct a spacetime as follows. Pick two timelike related points $q \ll p$ of Minkowski spacetime $\mathbb{R}^4$, η_{ab} and remove a closed ball K in the interior of $J^-(p) \cap J^+(q)$ (see Fig. 2.4). The resulting spacetime has a trivial singularity which can be doctored as followed. Choose a scalar function Ω which blows up rapidly as the (missing set) K is approached, and define a new metric by $g_{ab} =: \Omega^2 \eta_{ab}$. The resulting spacetime $\mathbb{R}^4 - K$, g_{ab} is inextendible and strongly causal. It is also non-singular in that there are no b-incomplete curves, no curvature blowup, no missing points, etc. But it fails to be globally hyperbolic. (This example also illustrates how the failure of Penrose's cosmic censorship leads to the possibility of supertasks discussed in chapter 4.) In sum, naked singularities in Penrose's sense form a third category of singularities conceptually distinct from the singularities associated with incompleteness and missing points.

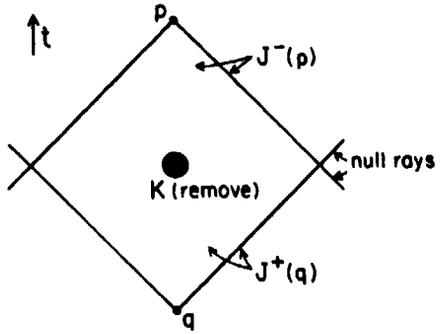

Fig. 2.4 A nakedly singular spacetime without singularities

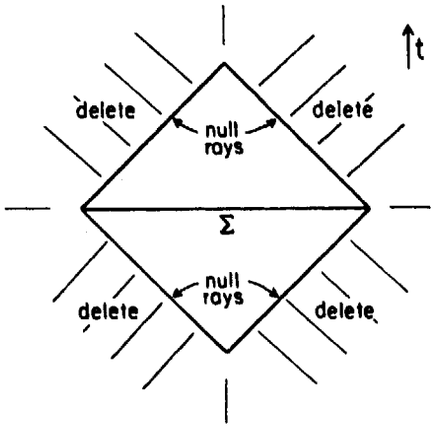

Fig. 2.5 A globally hyperbolic spacetime that is not maximal

The statement of cosmic censorship implicitly assumes that the spacetime model is maximal. The issue of maximality here is most easily discussed by using the fact that a globally hyperbolic spacetime contains a Cauchy surface. If M, g_{ab} has a Cauchy surface Σ, then there is a unique (up to isometry) maximal extension $\tilde{M}, \tilde{g}_{ab}$ of M, g_{ab} in which Σ is still a Cauchy surface (Choquet-Bruhat and Geroch 1969). The spacetime pictured in Fig. 2.5 is a diamond-shaped cutout from two-dimensional Minkowski spacetime. It cannot be extended in any way that keeps Σ (or any other Cauchy surface) Cauchy. But since it is extendible it does not satisfy the intended strong form of cosmic censorship. I will return to these matters in chapter 3.

There is also a connection between cosmic censorship and the recently discovered exotic differentiable structures $\mathbb{R}^4_E$ for topological $\mathbb{R}^4$. These are structures which are compatible with the topology of $\mathbb{R}^4$ but are not diffeomorphic to the standard atlas of charts.[25] Such exoticness cannot bloom

in the presence of Penrose's cosmic censorship hypothesis, at least if the Poincaré conjecture is correct.[26] If M, g_{ab} is a four-dimensional globally hyperbolic spacetime, then as follows from Dieckman (1988), M is diffeomorphically $\mathbb{R} \times \Sigma$, where Σ is a three-manifold. However, it follows from a theorem of McMillan (1961) that if the Poincaré conjecture is true, exotic $\mathbb{R}^4_E$ cannot be diffeomorphically such a product.

2.7 What is a spacetime singularity (again)?

Haven't we already answered this question to the best of our ability? To be honest, no. The main target of our quest is the set of essential singularities, that is, singularities that cannot be removed by extending the spacetime. (In the b-incompleteness classification these are singularities that correspond to non-regular points; in the Scott classification they are the non-removable singularities.) But this is not a well-defined set until the continuity/differentiability (c/d) conditions on the allowed extensions are specified. There are twin dangers to be avoided in making the specification. On the one hand we do not want to set the c/d conditions so high that they exclude extensions that might be realized in some physically reasonable situation. On the other hand, we do not want to set the c/d conditions so low that anything goes. At the extreme, if no c/d conditions are imposed then no spacetime will be counted as having an essential singularity (see Clarke 1993, pp. 115–116). For example, in the case of an FRW universe with a big bang and a big crunch we could "extend" both forward and backward in time and imagine an oscillating universe that recycles itself in a never-ending series of expansions and contractions. And indeed, such scenarios are to be found in the early scientific literature (see chapter 7) and in current popular science writings. But while it is amusing to construct such scenarios, one wonders whether they belong to art rather than science. Are such constructions constrained by physical principles, or is imagination the only guide?

At this juncture one might eschew the task of trying to specify a cut between essential and inessential singularities in favor of classifying singularities as to their strengths. Thus, for example, one could say that a spacetime has a C^k (for $k = 0^-, 0, 1^-, 1, \ldots$) singularity in the sense of b-incompleteness if there is no C^k extension through the singularity; the lower the k, the stronger the singularity. If, however, one tries to choose a cutoff point for essential singularities, a plausible starting position is to set the c/d condition on allowable extensions by what is necessary and sufficient to make sense of the laws of GTR, principally EFE.

At a minimum, one wants EFE to be defined in the sense of distributions.[27] Geroch and Traschen (1987) have argued that the appropriate class of metrics for meeting this requirement are the *regular metrics*: a symmetric tensor field g_{ab} is a regular metric just in case (i) the inverse g^{ab} exists and both g_{ab} and g^{ab} are locally bounded, and (ii) the weak first derivative of g_{ab} exists and is

locally square integrable.[28] A C^2 metric is regular, but much less well-behaved metrics are also regular. To understand the motivation behind this definition, note that if g_{ab} is a smooth metric, the curvature tensor for g_{ab} can be written as

$$R_{abc}{}^{d} = -2\Gamma^{d}{}_{m[a}\Gamma^{m}{}_{b]c} - 2\partial_{[a}\Gamma^{d}{}_{b]c} \tag{2.1}$$

where

$$\Gamma^{a}{}_{bc} = (1/2)g^{am}(2\partial_{(b}g_{c)m} - \partial_{m}g_{bc}) \tag{2.2}$$

The philosophy of the Geroch–Traschen approach is that if $R_{abc}{}^{d}$ is to be interpreted as a distribution, then each term on the right-hand side of the defining equation (2.1) should be independently interpretable as a distribution since one does not want to rely on the happenstance that the terms add up in just the right way to produce a well-defined distribution. That the metric is regular does guarantee that each term on the right-hand side of (2.1) is meaningful as a distribution. To see this, note first that if the derivatives of g_{ab} are interpreted as weak derivatives, then (2.2) is meaningful for a regular metric, and $\Gamma^{a}{}_{bc}$ will be locally square integrable. Thus, the first term on the right-hand side of (2.1), which is the product of two Γ's, will be a locally integrable function and thereby will define a distribution.[29] By the same token, $\Gamma^{a}{}_{bc}$ defines a distribution, and since the ordinary derivative of a distribution is a distribution, the second term on the right-hand side of (2.2) is also meaningful as a distribution. Some further calculation shows that taking the outer product of any number of g_{ab}'s or g^{ab}'s for a regular metric and the $R_{abc}{}^{d}$ from a regular metric results in a distribution. By taking contractions of such products, it follows that for a regular metric the Einstein tensor $G_{ab} =: R_{ab} - (1/2)g_{ab}R$ is meaningful as a distribution. So the vacuum EFE $G_{ab} = 0$ make distributional sense for a regular metric; and so do the full EFE $G_{ab} = 8\pi T_{ab}$, at least if $T_{ab} = (1/8\pi)G_{ab}$ is used as a definition of the energy–momentum tensor. In sum, the regularity of the metric is sufficient to guarantee that EFE are distributionally meaningful. There is no airtight argument to the effect that regularity is a necessary as well as sufficient condition; for after all, there are distributions that do not arise from locally integrable functions, the famous Dirac δ-function being a prime example. But it is hard to see how to weaken the requirement of regularity and still guarantee that G_{ab} is a distribution.

Requiring that the metric of the extended spacetime be regular may be necessary for a physically meaningful extension, but it arguably does not go far enough. For the satisfaction of EFE in a distributional sense does not guarantee that these equations attain their intended physical content. For instance, the *Bianchi identity* $\nabla_{[a}R_{bc]d}{}^{e} = 0$ need not make sense for a regular metric. But this identity is crucial to two intended implications of EFE. The first implication concerns conservation principles. It is the Bianchi identity that entails that $\nabla_a G^{ab} = 0$. The application of EFE then produces the local conservation law $\nabla_a T^{ab} = 0$. In the second place, the Bianchi identity is also crucial to the initial value formulation of GTR. EFE can be divided into

constraint equations, which constrain the initial data on a spacelike hypersurface, and the evolution equations, which specify how the initial data is to be evolved off the initial hypersurface. The Bianchi identity is used to prove that if the constraints are satisfied at the initial moment, then they will be preserved by the time evolution. Thus, from the point of view of physics, one wants enough c/d to insure that the expression $\nabla_{[a}R_{bc]d}{}^{e}$ makes sense. This seems to require that not only is the metric regular but also that the weak second derivative of the metric exists and is locally square integrable.[30]

One may also want to make sure that the intended deterministic development of the spacetime structure is secured. Thus, in the matter-free case ($G_{ab} = 0$) if the value of the metric and its time derivative are specified on a spacelike hypersurface Σ—technically, the first and second fundamental forms of Σ (see chapter 3)—then one wants to be able to prove the local existence and uniqueness of solutions to Einstein's vacuum field equations corresponding to the given initial data.[31] Fischer and Marsden (1979) stated the weakest known sufficient conditions for local existence and uniqueness in terms of Sobolev spaces of non-integer dimension. A sufficient condition for their sufficient condition is that the metric as restricted to the initial value surface be C^3 and the normal derivative of the metric be C^2. It is apparently not known what the minimally sufficient conditions are.

In sum, it is far from clear what c/d conditions should be required of physically meaningful extensions. Presumably, the metric should be at least regular, but what further conditions should be imposed remains up for grabs, although there are a number of obvious factors that go into the decision. Further discussion will have to be hypothetical, supposing one or another of the decisions that could be made. For sake of discussion then, suppose first that it has been decided that physically meaningful extensions require that the metric must be at least C^2. Then most of the familiar solutions to EFE with singularities—FRW, Kruskal–Schwarzschild, etc.—will be counted as essentially singular. Moreover, with the help of a little Leibnizian metaphysics we can apply the Hawking–Penrose singularity theorems (to be discussed in section 2.8) to conclude that essential singularities can be expected to occur under generic conditions in metaphysically acceptable models of gravitational collapse and cosmology. These theorems assume that (a) each $p \in M$ and each direction $\mathbf{V} \in M_p$ together define a unique geodesic, and (b) the geodesic depends differentiably on the initial data. Condition (a) requires that the metric be C^{2-} (i.e., the second derivative exists and is locally bounded), and (b) requires that the metric be C^2. But since a C^{2-} metric can be approximated by C^2 metrics, C^{2-} suffices for both (a) and (b). If we are given a C^2 spacetime M, g_{ab} to which the conditions of the Hawking–Penrose theorem apply, we may conclude that the spacetime is singular in the sense of geodesic incompleteness. Of course, this incompleteness may be due simply to the fact that the spacetime is extendible. We know from section 2.3 that every C^2 spacetime is extendible to a maximal C^2 spacetime. So either the spacetime in question is maximal or else not. In the former case the spacetime is

essentially singular. In the latter case the singularities may not be essential, but the spacetime can be ignored on the grounds that it violates Leibniz's principle of sufficient reason (see section 2.3).

It would be nice to be able to draw conclusions about essential singularities from existing singularity theorems. But there is a potential awkwardness to the situation we have envisioned. Impulse wave solutions (metric C^{1-}, i.e., the first derivative exists and is locally bounded) will be counted as essentially singular. More generally, solutions to EFE will be counted as essentially singular even though there are well-defined geometrical extensions of differentiability lower than C^2 (or C^{2-}). In this sense there would be nothing in the geometry of spacetime itself to prevent extending through the singularity, and only the failure of the physical laws (that by supposition require C^2 metrics) to keep up with the geometry would lead us to say that the singularity is essential. This parting of company of geometry and physics in a theory that was supposed to marry them is simply the price that has to be paid in the current approach to recognizing essential singularities.

Suppose next that we decide that the necessary and sufficient conditions c/d for physically meaningful extensions are weaker than C^2 (or C^{2-}). One could still talk about singularities in the sense of b-incompleteness if the metric were C^{1-}. One result of the present supposition would be that fewer spacetimes would be counted as essentially singular than on the previous supposition. Another result would be that the Hawking–Penrose theorem would no longer apply in its present form. It remains to be seen whether this and related theorems can be reconstituted so as to predict the occurrence of essential singularities under the hypothesized low c/d conditions. On the other hand a potential benefit is the possibility of restoring a concordance between geometry and physics by showing that for an essential singularity something closer to a breakdown in the spacetime geometry prevents an extension.

It may be hard to believe that anything physically significant depends on finicky details of the c/d conditions on the metric. But we have to be prepared for the awful possibility that Nature does not share our beliefs. In an optimistic vein, Hawking and Ellis wrote:

> In fact, the order of differentiability of the metric is probably not physically significant. Since one can never measure the metric exactly but only with some margin of error, one could never determine that there was an actual discontinuity in its derivatives of any order. Thus one can always represent one's measurements by a C^∞ metric. (Hawking and Ellis 1973, p. 58)

The danger of this sort of reasoning has been emphasized by Chruściel (1992). The silent premise in the Hawking–Ellis argument is that any C^k ($k \geq 0$) field on a C^∞ manifold can be approximated to arbitrarily great accuracy by a C^∞ field. Their conclusion is that for purposes of physics one might as well assume that the field is C^∞. By parallel reasoning one could argue that since a C^∞ field on a manifold can be approximated to any desired accuracy by an analytic field, we might as well assume that the field is analytic. But this

latter conclusion is misleading with respect to the problem of temporal evolution in GTR; for while an analytic function is determined by its values in any neighborhood, the non-analytic solutions of EFE do not have this property. Moreover, as Chruściel emphasizes, imposing differentiablity assumptions on evolved solutions may actually be incompatible with EFE. Hyperbolic partial differential equations are compatible with the loss of degrees of differentiability through time. So while it may be justified to stipulate that the initial conditions enjoy a certain smoothness, the stipulation that this smoothness holds for all times may clash with EFE. More generally, if one takes a non-instrumentalist view of spacetime structure, then continuity or differentiability of the spacetime metric is either there or not, regardless of whether we can directly detect it with our measuring instruments. And the differences can have important consequences for the interpretation of singularities.

The upshot of this discussion is rather unsettling. Since we are left in a state of uncertainty about what c/d conditions to require of extensions, we are left with corresponding uncertainties about what an essential spacetime singularity is and about how generic essential singularities are among the solutions to EFE. Nevertheless, we do know some of the necessary conditions that a physically meaningful extension should satisfy. This is enough to show in some cases that the singularities involved are essential, but the demonstration will be postponed until chapter 7.

Finally, a word should be said about the differentiable structure for the spacetime manifold, which is often taken as an unproblematic given. An illustration of how this structure may make a difference for singularities, at least at the level of mathematics, has come to light in connection with the exotic differentiable structures $\mathbb{R}^4_E$ for topological $\mathbb{R}^4$. It is known that flat metrics (of any signature) which are smooth with respect to $\mathbb{R}^4_E$ cannot be geodesically complete (see Brans and Randall 1993). So either the exotic $\mathbb{R}^4_E$ do not admit flat Lorentzian metrics, which would be an interesting result in itself; or else they do, and all these metrics are singular. In the latter case there arises the question of the basis of the incompleteness. For some $\mathbb{R}^4_E$'s there is no hope of completing such a metric by means of an extension into the standard $\mathbb{R}^4_S$ since some $\mathbb{R}^4_E$'s cannot be smoothly imbedded into $\mathbb{R}^4_S$; nor, of course, can the completion be accomplished by extending to a flat metric in $\mathbb{R}^4_E$. On the other hand, we know by the result of Clarke (1973) that *local* extensions are possible. There is no topological obstruction to a global extension, as with the quasi-regular singularity in Fig. 2.1 since we are dealing with topological $\mathbb{R}^4$. So is the incompleteness due to acausal behavior, as in Fig. 2.3, or is some new sort of differentiable obstruction involved?

2.8 Singularity theorems

How widespread are spacetime singularities among the solutions to EFE? One approach to answering this question would be to take a headcount of the

known exact solutions to EFE. But such a count can hardly be expected to give a reliable reflection of the situation in the class of all solutions since the solutions we have in hand are perforce of a simple and, therefore, special character. Another approach would be to attempt to establish theorems that characterize general conditions under which singularities will occur. There was no motivation to search for such theorems as long as it was believed that spacetime singularities were artifacts of the idealizations involved in constructing models of gravitational collapse and cosmology. Thus, it was not until the 1960s that much progress was made on the second approach. At first there were some false starts. For instance, in the second English edition of *The Classical Theory of Fields*, Landau and Lifshitz reached the "fundamental conclusion that the presence of a singularity in time is not a necessary property of cosmological models of the general theory of relativity, and that the general case of an arbitrary distribution of matter and gravitational field does not lead to the appearance of a singularity" (Landau and Lifshitz 1962, p. 397; see also Lifshitz and Khalatnikov 1963). Their argument was mistaken and was later withdrawn. From the mid-1960s onward advances were rapidly made, culminating in the main Hawking–Penrose theorem (1970). Since it is one of the most important results in modern GTR and is central to our concerns, it will be quoted and explained here.

Theorem (Hawking and Penrose). Let M, g_{ab} be a time-oriented spacetime satisfying the following four conditions:

(1) $R_{ab}V^aV^b \geq 0$ for any non-spacelike V^a.
(2) The timelike and null generic conditions are fulfilled.
(3) There is no closed timelike curve.
(4) At least one of the following holds:
  (a) There exists a compact achronal set without edge.
  (b) There exists a trapped surface.
  (c) There is a $p \in M$ such that the expansion of the future (or past) directed null geodesics through p becomes negative along each of the geodesics.

Then M, g_{ab} contains at least one incomplete timelike or null geodesic.

Several explanatory comments are in order. The theorem is purely geometrical—it uses no physical laws. Physics enters in justifying the various hypotheses of the theorem. If EFE (without cosmological constant) are satisfied and the energy–momentum tensor T^{ab} satisfies the *strong energy condition*: $T_{ab}\zeta^a\zeta^b \geq -(1/2)T$, $T =: T^a{}_a$, for any unit timelike ζ^a, then (1) will hold. The strong energy condition is believed to be valid for any physically reasonable classical source field. For a precise statement of the *generic conditions* (2) the reader is referred to Hawking and Ellis (1973). These conditions will be satisfied provided each timelike or null geodesic experiences tidal force at some point in its history. So far then, the presumptions of the

theorem seem innocuous. The meat of the presumptions comes in condition (4). Condition (4a) says that, at least at one time, the universe is closed and that the compact slice corresponding to this time is not intersected more than once by a future-directed timelike curve. Condition (4b) refers to a *trapped surface*. This is a compact, two-dimensional submanifold $\mathfrak{T} \subset M$ such that the expansion of both ingoing and outgoing future-directed null geodesics orthogonal to $\mathfrak{T}$ is everywhere negative. The physical significance of this condition derives from the fact that trapped surfaces are expected to form in gravitational collapse. Indeed, a result of Schoen and Yau (1983) shows that when a sufficient amount of matter is concentrated in a small enough region, a trapped surface results. Condition (4c) on the expansion of null geodesics can be expected to hold in various scenarios when the universe is collapsing in the past or future directions.

To convey some sense of how singularity theorems are proved, it will be helpful to concentrate on a simpler and less powerful theorem. Consider a spacetime that is globally hyperbolic (see section 2.5) and that, therefore, possesses a Cauchy surface Σ. Cauchy surfaces have a number of nice properties, but the only one needed here is that for any point p, say to the future of Σ, there is a longest timelike curve from p to Σ. This curve will be a geodesic γ (because in relativistic spacetimes timelike geodesics maximize rather than minimize proper length) and it will be orthogonal to Σ (because otherwise one could obtain a longer curve by veering off γ near where it meets Σ). Let us then focus on the congruence of timelike geodesics orthogonal to Σ. If V^a is the unit timelike vector field tangent to the congruence, the *expansion* θ of the congruence is defined by $\theta =: \nabla_a V^a$. The *Raychaudhuri equation* (a purely geometrical relation) shows that the rate of change $\dot{\theta} =: d\theta/d\tau = V^a \nabla_a \theta$ (where τ is proper time) is given by

$$\dot{\theta} = -\tfrac{1}{3}\theta^2 - \sigma_{ab}\sigma^{ab} + \omega_{ab}\omega^{ab} - R_{ad}V^aV^b \tag{2.3}$$

where σ_{ab} and ω_{ab} are respectively the *shear* and *rotation* matrixes for the congruence.[32] Since the congruence is orthogonal to Σ, ω_{ab} is initially zero; and since ω_{ab} vanishes at all times if it vanishes at any time, $\omega_{ab} = 0$. The second term on the right hand side is obviously non-negative. And by condition (1) the fourth term is also non-negative. Thus, we can conclude that

$$\dot{\theta} + \tfrac{1}{3}\theta^2 \leq 0 \tag{2.4}$$

Integrating (2.4) yields

$$\frac{1}{\theta(\tau)} \geq \frac{1}{\theta_0} + \tfrac{1}{3}\tau \tag{2.5}$$

It follows that if the initial expansion θ_0 is negative, that is, the congruence

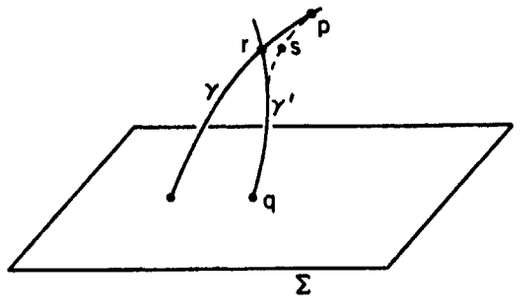

Fig. 2.6 An illustration of conjugate points

is converging, then $\theta \to -\infty$ (the geodesics begin to cross[33]) within a proper time $\tau \leq 3/|\theta_0|$. Now suppose for purposes of contradiction that there were a timelike curve from Σ into the future whose proper length exceeded $3/|\theta_0|$. Choose a point p on the curve lying a proper length greater than $3/|\theta_0|$ to the future of Σ. We know that there must be a longest timelike curve from p to Σ and that it will be a timelike geodesic γ orthogonal to Σ. But γ cannot be the longest curve. For a neighboring geodesic γ' will intersect γ at a point r between p and Σ (see Fig. 2.6). By constructing qsp, a timelike curve of proper length greater than $3/|\theta_0|$ can be obtained. To escape the contradiction we must conclude that no timelike curve from Σ into the future has a length greater than $3/|\theta_0|$ and, thus, that every timelike geodesic is incomplete.

This simple singularity theorem is limited to collapsing universes with Cauchy surfaces, but proofs of the more far-reaching theorems, though much more difficult, use many of the same concepts. The argument given above is an application of the concept of *conjugate points*. A point r on a geodesic γ belonging to a congruence of geodesics orthogonal to a spacelike Σ is said to be conjugate to Σ along γ if (intuitively) geodesics that are infinitesimally close to γ at Σ begin to cross at r. This will happen just in case (as in the illustration above) $\theta \to -\infty$ at r. More generally, consider any congruence of timelike geodesics, hypersurface orthogonal or not. Points p and q on some geodesic γ of the congruence are conjugate just in case infinitesimally nearby geodesics intersect γ at p and q. The importance of this concept lies in the fact that a timelike geodesic from p to q maximizes proper length between p and q if and only if there is no point conjugate to p between p and q. Conditions (1) and (2) of the Hawking–Penrose theorem are used to prove the existence of conjugate points on complete geodesics. Thus, in a spacetime satisfying (1) and (2), if all timelike geodesics are complete, they could not maximize proper length. A contradiction is obtained if other hypotheses can be shown to entail the existence of a longest timelike curve. The presence of a Cauchy surface will suffice, but so will much weaker conditions, although the proof that weaker conditions suffice is much more elaborate. Readers interested in the details may consult Hawking and Ellis (1973) and Wald (1984a).

At this juncture two features of the Hawking–Penrose theorem should be emphasized. The first is that, under the stated conditions, the theorem demonstrates the occurrence of singularities in the form of geodesic incompleteness and, a fortiori, b-incompleteness. But it gives no information about the type of singularity that can be expected to occur. I will have more to say about this matter in section 2.9.

The second feature which calls for comment is the causality condition (3). Such an assumption is harmless as long as "no time travel" is regarded as an a priori necessary condition for physically reasonable models of the universe. However, such an attitude is no longer as popular as it once was. As R.P.A.C. Newman notes:

> It has become customary to claim that closed timelike curves render a spacetime physically unreasonable. Certainly, if the universe does contain closed timelike curves, a revision of fundamental premises of physics, and philosophy, may be necessary. However, to dismiss this, and other forms of causality violation, out of hand is reminiscent of the dogmatism regarding singularities prior to the singularity theorems. (Newman 1989, p. 982)

There is a singularity theorem due to Hawking (1967) that does not explicitly mention causality.

Theorem (Hawking). Let M, g_{ab} be a time-oriented spacetime satisfying the following three conditions:

(1′) $R_{ab}V^aV^b \geq 0$ for every non-spacelike V^a.
(2′) There exists a compact spacelike hypersurface $\Sigma \subset M$ without edges.
(3′) The unit normals to Σ are everywhere converging (or diverging).

Then M, g_{ab} is timelike geodesically incomplete.

Condition (1′) is the same as (1) in the Hawking–Penrose theorem. Condition (3′) says that at the time corresponding to Σ the universe is everywhere contracting (or expanding). Although the Hawking theorem does not explicitly impose causality constraints, they are implicit in the hypotheses. For if the spacetime is simply connected, the time slice Σ postulated by (2′) cannot be intersected more than once by a future-directed timelike curve. Thus, (2′) rules out various acausal spacetimes; in particular, viciously causal ones such that a closed timelike curve passes through every point (at least assuming simple connectedness).

One is thus led to wonder whether violations of causality can prevent the occurrence of singularities. Newman (1989) has shown by explicit example that the chronology condition in the Hawking–Penrose theorem cannot be entirely removed. Starting from a conformal modification of a three-dimensional Gödel spacetime, he produced a four-dimensional spacetime that satisfies conditions (1), (2), has a (non-achronal) trapped surface, but is timelike and null geodesically complete. It is an open question whether

examples can be constructed with achronal trapped surfaces. If so, one can envision cases of gravitational collapse which form black holes, the interiors of which contain not singularities but causality violations.

Tipler (1976, 1977a) showed that in some restricted cases closed timelike curves will be accompanied by singularities in the form of null geodesic incompleteness. However, Tipler's results cannot be considered as a generalization of the Hawking–Penrose theorem since the crucial condition (4) of this theorem plays no role in his proofs. Kriele (1990a) has proved a direct generalization of the Hawking–Penrose theorem in which the chronology condition (3) is replaced by a weaker but more complicated causality condition which (very roughly) prohibits almost closed causal curves in the surface generated by the null geodesics that are orthogonal to the trapped surface. Kriele (1989, 1990b) also supplied a result that says in effect that chronology violations which "do not extend to infinity" are accompanied by singularities.

Theorem (Kriele). Let M, g_{ab} be a time-oriented spacetime such that

(1″) $R_{ab}K^aK^b \geq 0$ for all null vectors K^a.
(2″) The null generic condition is satisfied.
(3″) The chronology-violating region $V \subset M$ has compact closure but $M - V \neq \varnothing$.

Then if $V \neq \varnothing$, the boundary of V is generated by almost closed incomplete null geodesics.

If one imagines that the chronology violation is manufactured by a time machine, one would like to strengthen this result to say that chronology violations that do not *start* at infinity necessarily involve singularities. This and other aspects of chronology violation will be taken up in chapter 6.

Summarizing the present situation, it seems fair to say that while existing results are far from definitive, they strongly suggest that violation of the chronology condition is not a promising way to avoid singularities.

Returning finally to the question with which this section began, are we any wiser about how widespread singularities are among the solutions to EFE? The question becomes more tractable if cases are divided. At a minimum we should distinguish between solutions with $\Lambda < 0$, $\Lambda = 0$, and $\Lambda > 0$, the expectation being that singularities are more prevalent for $\Lambda = 0$ than for $\Lambda > 0$ and more prevalent still for $\Lambda < 0$. Cases can be further subdivided between spatially open and spatially closed universes, the expectation being that singularities are more prevalent in the latter cases. A further subdivision would distinguish between vacuum solutions ($T^{ab} = 0$) and non-vacuum solutions ($T^{ab} \neq 0$), the expectation being that singularities are more prevalent when matter–energy is present. For some sub-sub-sub- . . . cases we may be able to prove unqualified results: every solution belonging to the said class is singular (at least in the sense of geodesic completeness). In other sub-sub-

sub- . . . cases singularities may not be universal but nevertheless generic. To make this precise one would have to define an appropriate topology on the space of solutions to EFE and show that within the chosen subclass the singular solutions form an open dense subset. I know of no such results in the literature; indeed, how to define the appropriate topology seems to be an open question (see Geroch 1970a). Nevertheless, intuition suggests that such results should be forthcoming, for the existing singularity theorems encourage the notion that when a solution to EFE is provably singular, a small perturbation of the solution will produce a spacetime that is also singular.

2.9 Singularities and quantum effects

Faced with results indicating that singularities are a generic feature of general relativistic spacetimes, two opposing attitudes are possible. One is that GTR forces us to recognize spacetime singularities as a new feature of reality and that we simply have to learn how to live with this new feature. The cosmic censorship hypothesis, to be studied in detail in chapter 3, is an attempt to make the coexistence a peaceful one. The opposing attitude is that no cohabitation is possible since spacetime singularities are absurdities. This second attitude promotes a reading of the Hawking–Penrose theorem and allied results as threatening a reductio of GTR or at least as indicating that GTR must break down under the conditions where the theory says that singularities will develop. To be more than wishful thinking this second attitude must be accompanied by some reason for believing that a more complete and adequate theory will retain the successes of GTR while showing how singularities are avoided. Some theoretical physicists espouse the hope that a quantum theory of gravity will fit the bill. Since the outlines of such a theory can only be dimly perceived, this hope cannot be assessed with any accuracy at the present time. Nevertheless, it is a useful exercise to try to lay out some of the thinking behind the hope.

To begin, why should one expect that quantum effects will come into play in the regimes where the Hawking–Penrose singularity theorems and allied results say that spacetime singularities develop? The answer would have to come in two parts. First, one would have to cite considerations from various approaches to quantum gravity to support the conclusion that when gravitational fields become sufficiently strong on some quantum scale, quantum effects will out. Second, one would have to find a way around the fact, noted in the discussion in section 2.3, that it is mathematically possible to have spacetime singularities (in the sense of geodesic incompleteness) that do not involve arbitrarily strong gravitational fields in the form of unbounded curvature. Thus, one would need to find support for what Clarke (1988) calls the *curvature conjecture*; namely, that under physically reasonable conditions, the singularities predicted by the Hawking–Penrose theorem and allied results will involve unbounded curvature. As an example of a result that lends some

support to this conjecture, Clarke (1975a, b) showed that in a C^{0-} inextendible globally hyperbolic spacetime that is not too specialized (in a technical sense), a singularity $p \in \partial_b M$ that lies at the end of a b-incomplete timelike curve is a curvature singularity. Combined with Penrose's form of cosmic censorship the support lent to the curvature conjecture would be very strong; but if cosmic censorship fails, the fate of the conjecture is left in doubt (see Clarke and Schmidt 1977; Clarke 1988 for further discussion).

The next question to be addressed is why one should think that quantum effects, assuming that they do kick in under the conditions that classical GTR says that spacetime singularities form, should rescue us from singularities. The first glimmering of hope has nothing to do with quantum gravity per se but simply with the amazing ability of quantum mechanics to smooth away classical singularities. As an example, compare Newtonian mechanics with ordinary quantum mechanics. For the former, singularities for point mass particles interacting via a $1/r^2$ force can develop either because the particles collide in a way that solutions to Newton's equations of motion cannot be continued, or else because the particles, without colliding, all disappear to spatial infinity in a finite amount of time (see Earman 1986). In the quantum mechanical treatment of this problem the Hamiltonian operator is (essentially) self-adjoint so that the time evolution operator, which is the exponentiation of the Hamiltonian, is unitary and is defined for *all* time $-\infty < t < +\infty$—the solution never breaks down!

More substance is given to this glimmering by two further considerations. The first concerns the treatment of the big bang singularity in the semiclassical approach to quantum gravity where there is no attempt to quantize the metric itself but where the effect of quantum fields on the spacetime geometry is calculated by computing the expectation value $\langle T_{ab} \rangle$ of the (renormalized) energy–momentum tensor of the fields and inserting the result back into EFE. In a generic early universe one would expect the spacetime geometry to be inhomogeneous and anisotropic. In some cases these lumpy conditions lead to particle creation, and when the back reaction on the metric is calculated the effect can be to remove the big bang as a genuine singularity. But depending on the details, the effect can also be to leave the big bang singularity while removing the particle horizons associated with it (see Anderson 1983, 1984 and also chapter 5 below).

The second consideration starts from the observation that the applicability of all of the aforementioned singularity theorems relies on various energy conditions on classical fields and that even the *weak energy condition* can be violated for quantum fields. (This condition requires that $T^{ab}V^aV^b \geq 0$ for all timelike V^a, which implies that the energy density as measured by any observer is non-negative.) A more positive note is sounded by Parker and Fulling (1973) who studied a closed FRW model filled with a neutral scalar quantum field possessing mass. They showed that in this regime coherent quantum states can give rise to negative pressures sufficiently large that $\langle T_{ab} \rangle$ violates the weak energy condition. As a result, instead of collapsing to a

singularity, the universe 'bounces' and reexpands. However, it should be noted that in some circumstances averaged or integrated versions of the classical energy conditions do hold for quantum fields[34] and that these averaged energy conditions may suffice for proving singularity results (see Tipler 1978).

So far I have been speaking of spacetime singularities in the sense relevant to the standard singularity theorems, that is, geodesic incompleteness or some generalized kind of incompleteness such as g.a.p. incompleteness. Naked singularities in the sense of violations of cosmic censorship (see section 2.6) may or may not involve incompleteness and/or curvature blowup. This raises the question of whether there is any hope that quantum gravity will help to censor such naked singularities. In chapter 6 I will mention results that indicate that quantum gravity may help to prevent the kind of naked singularities involved in the manufacture of closed timelike curves. But the mechanism of censorsip involves the divergence of $\langle T_{ab} \rangle$ and an attendant curvature blowup. On the other side of the ledger it should be noted that quantum considerations can make things much worse with respect to naked singularities. Suppose, in accord with one popular version of the cosmic censorship conjecture, that the only singularities that develop in physically reasonable cases of gravitational collapse in classical general relativity are those that are safely hidden inside of black holes (see chapter 3 for details). Hawking radiation, a quantum field theoretic effect, leads to the eventual evaporation of the black hole. And insofar as classical spacetime geometry can be used to describe the processes, a naked singularity will result from the evaporation (see chapter 3). Of course, if quantum gravity manages to avoid singularities, then what will be visible to observers when a black hole evaporates is not literally a spacetime singularity but a region where quantum effects dominate. The attendant quantum uncertainties associated with such effects may be just as disruptive to predictability and determinism as are the naked singularities of classical GTR.[35]

Such intimations about the relation between quantum effects and spacetime singularities are solid in the sense that they are based on well-established techniques of semiclassical quantum gravity. But precisely because of their semiclassical character they may not provide reliable indications of what happens if, as many theorists believe, the spacetime metric itself must be quantized. There are various speculations about what will happen, the most widely publicized being Hawking's (1988) proposal for a "no boundary condition" for quantum cosmology, a condition that banishes the initial singularity implied by classical general relativity for big bang universes. At present, however, there are no clear experimental or theoretical guidelines for evaluating such speculations.[36]

2.10 Conclusion

The above probing of the idea of spacetime singularities has revealed four distinct root concepts: singularities as involving a breakdown of the metric

because, for example, of a blowup of the curvature at a finite distance; singularities as involving geodesic incompleteness or some more general sense of incompleteness such as generalized affine parameter incompleteness; singularities as involving missing points; and singularities as involving violation of cosmic censorship (naked singularities). The above discussion also helped to reveal that within each of the first three categories there is a sizable family of subcategories. The discussion in chapter 3 will show that 'naked singularity' is a name for an extended family of notions. In short, spacetime singularities exhibit a richness and complexity unimagined by the early pioneers of GTR. The more-or-less official definition of singularities adopted in the physics literature is in terms of incompleteness. This choice, quite frankly, seems to have been guided by expediency: this is the sense that most easily lends itself to proofs of the existence of singularities.

Our knowledge of the prevalence of singularities among the solutions of EFE is rudimentary but growing. The received wisdom is that the Hawking–Penrose theorem and allied results establish that singularities in the sense of geodesic incompleteness are endemic in general relativistic spacetimes. That conclusion is warranted if essential singularities are defined in terms of extensions whose metrics are at least C^{2-}. Whether the conclusion holds up under substantially weaker differentiability requirements is unknown. There are some results connecting geodesic incompleteness to singularities in some more intuitive sense such as curvature blowup; but the connection is still not well understood in general. And as chapter 3 will show, there are few informative results about the occurrence and non-occurrence of naked singularities.

Because of the absence of a technique that yields intuitively satisfactory results for attaching singular points to the spacetime manifold, it is perhaps best to drop talk of spacetime singularities—which suggests localizable objects—in favor of talk about singular spacetimes—which does not carry any such suggestion. However, talk about spacetime singularities is too well entrenched to fight.

In many ways we are still ignorant about spacetime singularities in classical GTR. And yet compared to what was known only a few decades ago, we know quite a lot about singularities. We certainly know enough to begin investigating the implications of singularities for classical general relativistic physics and spacetime philosophy. This will be the task undertaken in the coming chapters. In advance let me lay some of my cards on the table. If asked to say briefly what is wrong with a singular spacetime, my short answer would be: nothing per se. Contrary to Einstein, I do not think the fact that GTR predicts spacetime singularities is necessarily a cause for alarm, and I certainly not think the prediction of singularities is a signal that the theory self-destructs. But there are singularities and there are singularities. Some types are associated with acausal features and/or a breakdown in determinism. So while spacetime singularities per se may not be objectionable, some of their attendant features do pose troublesome questions for physics. These questions deserve detailed attention. And this book aims to provide it.

Notes

1. Szekeres noted that Synge (1950) had proposed a definition of spacetime singularities but that this definition could hardly be regarded as satisfactory since it depended on the choice of coordinate system.

2. A survey of the state of the art circa 1970 is given in Geroch (1970a). Tipler et al. (1980) summarized the situation to that point. I am not aware of any comparable review article that brings things up to the present. But the book by Clarke (1993) contains an authoritative statement of our present understanding of many technical issues about singularities.

3. It will be assumed throughout that M is a C^∞ manifold. This means that the transformations between two charts or coordinate systems is infinitely continuously differentiable; see Wald (1984a, Ch. 2) for details. If one starts with a C^k $(k \geq 1)$ manifold, where the coordinate transformations are only C^k for some $k < \infty$, one can always produce a C^∞ atlas by winnowing down the coordinate charts in the original atlas. A *Lorentz metric* g_{ab} is a non-degenerate tensor field of type (0, 2) and signature $n - 1$ where $n = \dim(M)$. (The *signature* of g_{ab} is the number of positive eigenvalues of g_{ab}.) In the main intended case of $n = 4$ (three space dimensions and one time dimension), the signature convention $(+++-)$ will be used. The metric is C^k just in case partial derivatives of order k exist and are continuous. The metric is C^{k-} just in case partial derivatives of order k exist and are locally bounded. Later in the discussion more complicated continuity/differentiability conditions on the metric will be discussed.

4. A tensor of type (m, n) is one with m contravariant and n covariant indices. The *outer product* of a tensor T of type (m, n) and a tensor T' of type (m', n') is a tensor $T \otimes T'$ of type $(m + m', n + n')$ such that $T \otimes T'(w_1, w_2, \ldots, w_{m+m'}; v^1, v^2, \ldots, v^{n+n'}) = T(w_1, \ldots, w_m; v^1, \ldots, v^n) \cdot T'(w_{m+1}, \ldots, w_{m+m'}; v^{n+1}, \ldots, v^{n+n'})$, where the v's are vectors (aka contravectors) and the w's are dual vectors (aka covectors). In the text the outer product is indicated simply by juxtaposing the two tensors, e.g., the outer product of R_{ab} and R^{cd} is denoted by $R_{ab}R^{cd}$.

5. In this case the metric will not be C^2, but the curvature tensor will still make sense as a distribution; see section 2.7.

6. At each point on $\gamma(v)$, where v is some parametrization of γ, the $e_i^a(v)$ form a basis of the tangent space $M_{\gamma(v)}$. $g_{ab}(\gamma(v))e_i^a(v)e_j^a(v) = 0$ if $i \neq j$, and 1 or -1 if $i = j$ according as $e_i^a(v)$ is spacelike or timelike.

7. Given a metric g_{ab}, there is a unique derivative operator ∇_a which is compatible with the metric in the sense that $\nabla_a g_{bc} = 0$. The action of ∇_a on a vector field V^b is given by $\nabla_a V^b = \partial_a V^b + \Gamma^b{}_{ac} V^c$ where ∂_a is the ordinary derivative operator and the *Christoffel symbol* Γ is defined by $\Gamma^b{}_{ac} =: (1/2)g^{bm}\{\partial_a g_{cm} + \partial_c g_{am} - \partial_m g_{ac}\}$. If T^a is the tangent to a curve γ, a vector field V^b along γ is said to be *parallely transported* just in case $T^a \nabla_a V^b = 0$.

8. Either the physical components of the curvature blowup in all p.p. frames or they blow up in none, for two such frames are related by a fixed Lorentz transformation.

9. A *geodesic* may be defined as a curve whose tangent T^a is parallely propagated, i.e., $T^a \nabla_a T^b = 0$ (see Wald 1984a, p. 41). A parametrization λ of the geodesic is said to be *affine* just in case $T^a = (\partial/\partial\lambda)^a$ satisfies this equation. Alternatively, a geodesic could be defined as a curve $\gamma(v)$ such that $T^a = (\partial/\partial v)^a$ satisfies $T^a \nabla_a T^b = f(v) T^b$, i.e., the tangent vector remains proportional to itself under parallel transport. But it is always possible to shift to a new parameterization which is affine. The affine parameters for a geodesic are unique up to a positive linear transformation $\lambda \to a\lambda + b$, $a > 0$. Thus, the notion of a geodesic of finite (or infinite) affine length is well defined. For a timelike geodesic, proper time $\tau = \int \sqrt{-ds^2}$ is an affine parameter.

10. This definition is redundant. Consider a curve $\gamma: [0, a), [0, a) \subset \mathbb{R}$. A point $p \notin \gamma([0, a))$ is counted as an *endpoint* of the curve if every neighborhood of p intersects the image $\gamma([0, a))$ of the curve in M. Such an endpoint that is not on the curve exists if and only if the curve is extendible in M in the direction away from $\gamma(0)$. In what follows I will abuse the distinction between a curve and its image in M.

11. See chapter 4 for examples. If two points of a relativistic spacetime can be joined by a timelike curve, then the *longest* such curve (if one exists) is a timelike geodesic. There is no shortest timelike curve joining the two points since for any $\varepsilon > 0$ one can construct a timelike curve which joins the points and which accelerates in a frantic enough way that its proper length is less than ε.

12. The symbol $\mathbb{R}^4$ will be used ambiguously to denote the topological space and the differentiable manifold obtained by adding the standard differentiable structure. Non-standard differentiable structures for $\mathbb{R}^4$ will be discussed below in sections 2.6 and 2.7.

13. A *diffeomorphism* $\varphi: M \to \tilde{M}$ between the C^∞ manifolds is a one-one C^∞ map. φ^* denotes the *carry along* by φ. For example, the carry along for tangent vectors $\varphi^*: M_p \to \tilde{M}_{\varphi(p)}$ is defined by the condition $(\varphi^* V)(f) = V(f \circ \varphi)$ for $V \in M_p$ and $f: \tilde{M} \to \mathbb{R}$ a C^∞ function. The idea is easily generalized to other geometric objects (see Wald 1984a, Appendix C). In the case where φ maps M into itself and the action of φ can be represented as carrying a point p with coordinates x^i to a new point $\varphi(p)$ with coordinates $x'^i(\varphi(p)) = x^i(p)$, the carry along of a geometric object is given by the rule that in the new coordinates the components of the new object at the new point are the same as the components of the old object in the old coordinates at the old point.

14. We will see in the following section, however, that there is a sense in which the cone singularity admits of local extensions.

15. The relation $\geq$ is said to be a *partial order* on the set S iff for any $a, b, c \in S$, (i) $a \geq a$, (ii) $a \geq b$ and $b \geq c$ imply $a \geq c$, and (iii) $a \geq b$ and $b \geq a$ imply that $a = b$. $T \subset S$ is *totally ordered* iff $\geq$ holds between all pairs of elements of T. $b \in S$ is said to be an *upper bound* for T iff $b \geq a$ for all $a \in T$. Finally, $m \in S$ is a *maximal element* for S iff for any $c \in S$, $c \geq m$ implies $m = c$. *Zorn's lemma*, an equivalent of the axiom of choice, asserts that if every totally ordered subset of S has an upper bound, then there is a (not necessarily unique) maximal element. The obvious idea for proving the existence of a maximal extension would be to take $(M', g'_{ab}) \geq (M, g_{ab})$ to mean that there is an isometric imbedding of the latter into the former. Start with a spacetime M, g_{ab} for which we would like to prove the existence of a maximal extension. Take S to be the set of all spacetimes M', g'_{ab} such that $(M', g'_{ab}) \geq (M, g_{ab})$. To prove the existence of an upper bound for any totally ordered subset T of S we can apply the inductive limit construction in which the union of spacetimes in T is taken and each spacetime in T is identified with its isometric image in the union (see Hawking and Ellis 1973, p. 249). Then if Zorn's lemma could be applied, the existence of a (not necessarily unique) maximal extension of M, g_{ab} would be immediate. Unfortunately the lemma cannot be applied because the relation of isometric imbeddability as defined above is not necessarily an order relation; in particular, property (iii) can fail: there are spacetimes such that $(M, g_{ab}) \geq (M', g'_{ab})$ and $(M', g'_{ab}) \geq (M, g_{ab})$ even though $(M, g_{ab}) \neq (M', g'_{ab})$ (i.e., the spacetimes are not isometric). For example, let M, g_{ab}

be the result of truncating Minkowski spacetime by removing all the points on or above a scallop-shaped time slice, and let M', g'_{ab} be constructed similarly but using a differently shaped time slice. The concept of a *framed spacetime* (see below) overcomes this difficulty; for the imbedding of M, g_{ab}, F into M', g'_{ab}, F', if it exists, is unique (see Geroch 1969).

16. That is, h_{ab} is of signature $n = \dim(M)$.

17. The terms b-completeness and b-incompleteness derive from the fact that they arise from Schmidt's (1971) bundle boundary construction; see below.

18. For details the reader is referred to Hawking and Ellis (1973) and Clarke (1993).

19. Roughly, the natural extension of the exponential map in M to $\hat{M}$ is required to be continuous; see Geroch et al. (1982) for details. (Assuming that the metric is C^2, for each $p \in M$ and each direction $T^a \in M_p$, there is a unique geodesic through p in the direction T^a. The *exponential map* is the map from the tangent space M_p to M which takes $T^a \in M_p$ to the point lying at a unit affine distance along the geodesic which passes through p in the direction T^a.) A different characterization of spacetime singularities, also subject to the difficulty under discussion, was given by Geroch (1968b).

20. Compare this to the treatment of Hawking and Ellis (1973, pp. 291–292).

21. When $\dim(M) \geq 2$, the *Weyl tensor* C_{abcd} is defined by

$$C_{abcd} = R_{abcd} - (2/(n-2))(g_{a[c}R_{d]b} - g_{b[c}R_{d]a}) - (2/((n-1)(n-2)))Rg_{a[c}g_{d]b}.$$

22. The following definition is suggested by ideas of Misner (1963) and Ryan and Shepley (1975), modified by using b-completeness in place of geodesic incompleteness.

23. Scott (1992) does not claim that her approach captures the missing points idea; that is my interpretation. What follows is a simplified and somewhat distorted version of Scott's analysis.

24. Let $\gamma: I \to M$, $I = [a, b) \subset \mathbf{R}$, be a parameterized half-curve, and let $\tilde{M}, \varphi$ be an envelopment of M. $p \in \tilde{M}$ is said to be a *limit point* of $\gamma(\lambda)$ just in case there is an increasing sequence of numbers $\lambda_i \in [a, b)$, $i = 1, 2, 3, \ldots$, such that $\lambda_i \to b^-$ and $(\varphi \circ \gamma)(\lambda_i) \to p$. p is said to be an *endpoint* just in case $(\varphi \circ \gamma)(\lambda_i) \to p$ for every increasing sequence λ_i such that $\lambda_i \to b^-$. I presume that boundary points which are limit points of half-curves of finite g.a.l. will also be endpoints of such curves. But such is not always the case for limit points of incomplete geodesics.

25. In other words, the global topological coordinates x, y, z, t for $\mathbf{R}^4$ are not everywhere smooth with respect to the coordinate charts of $\mathbf{R}^4_E$.

26. I am grateful to Professor Robert Gompf for clarifying this point. The Poincaré conjecture, which remains unsettled, asserts that a three-manifold is topologically S^3 if it is compact and simply connected.

27. Einstein's field equations are partial differential equations involving second-order derivatives of the metric tensor. The theory of distributions allows these equations to make sense in some cases where the metric fails to be twice differentiable. Technically, a *distribution* is a linear functional on a normed space of smooth functions. See Choquet-Bruhat et al. (1982) for a treatment of distributions in physics.

28. Choose a volume element dV on $\mathbf{R}^4$, say, the usual d^4x. The functions $f_{,a}$ are the *weak derivatives* of a function $f: \mathbf{R}^n \to \mathbf{R}$ iff for any C^∞ function ψ on $\mathbf{R}^n$ with compact support ("test function"), $\int f_{,a}\psi \, dV = -\int f(\partial\psi/\partial x^a) \, dV$. $f: \mathbf{R}^n \to \mathbf{R}$ is *locally square integrable* just in case $\int |f|^2 \, dV < \infty$ for any compact domain of integration in the support of the function. These ideas have natural generalizations to arbitrary n-dimensional manifolds.

29. If $f: \mathbf{R}^n \to \mathbf{R}$ is a locally integrable function with respect to the volume element dV, it defines a distribution D_f whose action on a test function ψ is given by $D_f(\psi) = \int f\psi \, dV$. This idea can be generalized to tensor fields on manifolds; see Choquet-Bruhat et al. (1982).

30. $\nabla_{[a}R_{bc]d}{}^e = 0$ involves the ordinary derivative of $R_{abc}{}^d$, which will be a distribution if $R_{abc}{}^d$ is a distribution, and the product of $\Gamma^a{}_{bc}$ and $R_{abc}{}^e$, which may fail to be a distribution even if $\Gamma^a{}_{bc}$ and $R_{abc}{}^d$ are distributions. But if $\Gamma^a{}_{bc}$ and $R_{abc}{}^d$ are each locally square integrable the product will be locally integrable and, therefore, will define a distribution.

31. Global existence and uniqueness is connected with the problem of cosmic censorship; see chapter 3.

32. If V^a is the normalized ($V^aV_a = -1$) tangent field of a geodesic congruence, then the associated *spatial metric* is $h_{ab} =: g_{ab} + V_aV_b$. The *shear* σ_{ab} and the *rotation* or *twist* ω_{ab} of the congruence are defined by $\sigma_{ab} =: \nabla_{(b}V_{a)} - (1/3)\theta h_{ab}$ and $\omega_{ab} =: \nabla_{[b}V_{a]}$. Rotation will be discussed further in chapter 6 in connection with the Gödel solution to EFE.

33. Strictly speaking, $\theta \to -\infty$ does not mean that the geodesics actually cross. The technically correct statement is that $\theta \to -\infty$ is the necessary and sufficient condition for the vanishing of a *Jacobi field* η^a along a geodesic γ of the congruence. η^a can be thought of as pointing from γ to an "infinitesimally close" geodesic of the congruence. See Wald (1984, Sec. 9.3).

34. The weak energy condition entails the *null energy condition* which requires that $T_{ab}K^aK^b \geq 0$ for every null vector K^a. Wald and Yurtsever (1991) showed that in two-dimensional curved spacetimes quantum fields obey an averaged version of the null energy condition. But they also showed that in four-dimensional spacetimes even this averaged version can fail.

35. As noted by Joshi (1993, p. 265). Just how naked singularities lead to disruptions of predictability and determinism will be discussed in chapter 3.

36. Another approach to combining general relativity and QM is provided by string theory. It has been claimed that this approach holds the promise of avoiding or rendering harmless spacetime singularities; see Kostelecký and Perry (1994).

3

Cosmic Censorship

It is one of the little ironies of our times that while the layman was being indoctrinated with the stereotypic image of black holes as the ultimate cookie monsters, the professionals have been swinging round to the almost directly opposing view that black holes, like growing old, are really not so bad when you consider the alternative.

Werner Israel (1986)

3.1 Introduction

The idea of cosmic censorship was introduced over twenty years ago by Roger Penrose (1969). About a decade later Penrose noted that it was not then known "whether some form of cosmic censorship principle is actually a consequence of general relativity." To which he added: "In fact, we may regard this as possibly the most important unsolved problem of classical general relativity theory" (Penrose 1978, p. 230). This sentiment has been echoed by Stephen Hawking (1979, p. 1047), Werner Israel (1984, p. 1049), Robert Wald (1984a, p. 303), Frank Tipler (1985, p. 499), Douglas Eardley (1987, p. 229), Stuart Shapiro and Saul Teukolsky (1991b, p. 330), Pankay Joshi (1993, p. 204), and many others. Thus, if an "important problem" in physics is one which is deemed to be important by leading research workers in the field, then the problem of cosmic censorship is undoubtedly near the top of the list for classical GTR. One of my goals here is to show why it is important in a more substantive sense. I also want to indicate why it is that, despite the intense effort that has been devoted to this problem, it remains unsolved. Indeed, the very statement of the problem remains open to debate.

A study of this topic can lead to payoffs in several areas of philosophy of science, two of which I will mention and one of which I will actually pursue. In my *Primer on Determinism* (Earman 1986) I attempted to deflate the popular image of determinism as unproblematically at work outside of the non-quantum domain. My message fell largely on deaf ears. But the

failure of cosmic censorship could well herald a breakdown in classical predictability and determinism of such proportions that it cannot be ignored.[1] Second, the growing band of philosophers of science who are turning towards an increasingly sociological stance will find the history of the cosmic censorship hypothesis a fascinating case study in the dynamics of a research program. Particularly interesting to me is how, in a subject with few hard results, the intuitions and pronouncements of a small number of people have shaped and directed the research. I leave the investigation of such matters to more capable hands.

3.2 Cozying up to singularities

As was seen in chapter 2, prior to the 1960s spacetime singularities were regarded as a minor embarrassment for GTR. They constituted an embarrassment because it was thought by Einstein and others that a true singularity, a singularity in the very fabric of spacetime itself, was an absurdity. But the embarrassment seemed to be a minor one that could be swept under the rug; for the then known models of GTR containing singularities all embodied very special and physically unrealistic features. Two developments forced a major shift in attitude. First, the observation of the cosmic low temperature blackbody radiation lent credence to the notion that our universe originated in a big bang singularity. Second, and even more importantly, a series of theorems due principally to Stephen Hawking and Roger Penrose indicated that, according to GTR, singularities cannot be relegated to the distant past because under quite general conditions they can be expected to occur both in cosmology and in the gravitational collapse of stars (see chapter 2). Thus, singularities cannot be swept under the rug; they are, so to speak, woven into the pattern of the rug. Of course, these theorems might have been taken as turning what was initially only a minor embarrassment into a major scandal. Instead, what occurred in some quarters was a 180° reorientation in point of view: singularities were no longer to be relegated to obscurity; rather, they were to be recognized as a central feature of the GTR, a feature which called attention to a new aspect of reality that was neglected in all previous physical theories, Newtonian and special relativistic alike. And thus we can hope to get definitive confirmation of GTR by confirming the presence of these new objects.

But before getting carried away with this newfound enthusiasm for singularities, we should pause to contemplate a potential disaster. If the singularities that occur in Nature are naked, then chaos would seem to threaten. Since the spacetime structure breaks down at singularities and since (pace Kant) physical laws presuppose space and time, it would seem that these naked singularities are sources of lawlessness. The worry is illustrated in Fig. 3.1 where all sorts of nasty things—TV sets showing Nixon's "Checkers

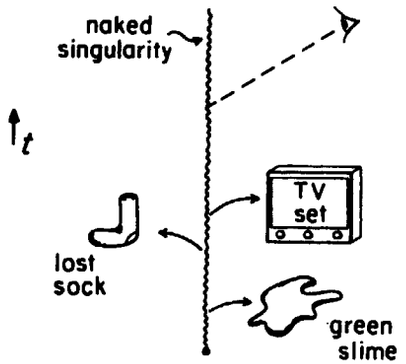

Fig. 3.1 A naked singularity disgorges

speech," green slime, Japanese horror movie monsters, etc.—emerge helter-skelter from the singularity.

The point can be put more formally in terms of the breakdown in predictability and determinism. If Σ is an achronal spacelike surface of a spacetime M, g_{ab} the *future* (respectively, *past*) *domain of dependence* $D^+(\Sigma)$ ($D^-(\Sigma)$) of Σ is defined to be the collection of all points $p \in M$ such that every causal curve which passes through p and which has no past (respectively, future) endpoint meets Σ. If $p \notin D^+(\Sigma)$ (respectively, $p \notin D^-(\Sigma)$) then it would seem that no amount of initial data on Σ will suffice for a sure prediction (respectively, retrodiction) of events at p since there are possible causal influences which can affect events at p but which do not register on Σ.[2]

To illustrate how naked singularities can lead to a breakdown in predictability and determinism, start with Minkowski spacetime $\mathbb{R}^4, \eta_{ab}$ and consider the spacelike hypersurface Σ corresponding to the level surface $t = 0$ of some inertial time coordinate t. $D^+(\Sigma)$ encompasses the entire future of Σ. Perform the now familiar trick of removing from $\mathbb{R}^4$ a closed ball K on the future side of Σ. The resulting spacetime has a naked singularity, the presence of which excludes the shaded region of Fig. 3.2 from $D^+(\Sigma)$. The future boundary of $D^+(\Sigma)$, called the *future Cauchy horizon* of Σ, is labeled as $H^+(\Sigma)$.[3] This naked singularity is rather trivial since it can be removed by extending the surgically mutilated spacetime back to full Minkowski spacetime. To make the example less trivial, one can choose a scalar field Ω that goes rapidly to zero as the missing region K is approached. The new conformally related spacetime $\bar{M}, \bar{g}_{ab}$, where $\bar{M} = \mathbb{R}^4 - K$ and $\bar{g}_{ab} = \Omega^2 \eta_{ab}$, is properly inextendible and so its naked singularity is irremovable.

Cosmic censorship is an attempt to have one's cake and eat it too. The idea is that we can cozy up to singularities without fear of being infected by the ghastly pathologies of naked singularities since GTR implies that, under reasonable conditions, Nature exercises modesty and presents us only with

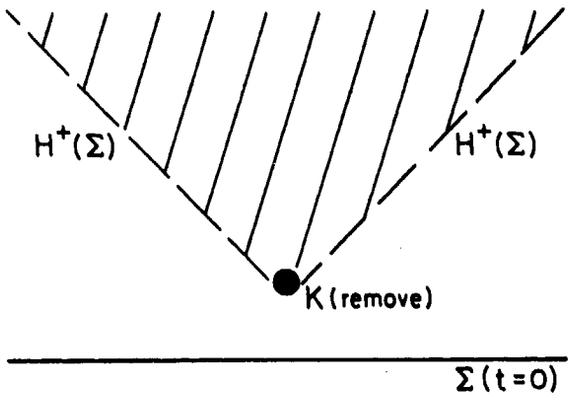

Fig. 3.2 How to create a naked singularity

singularities that have been clothed in some appropriate sense. The task of the following section is to try to understand in more precise terms what this means.

Before turning to this task, I should emphasize that while the approaches to cosmic censorship I will explore are motivated largely by concerns with predictability and determinism, there are a number of other reasons why physicists would like to believe that a cosmic censor is at work. I will mention three. First, if a suitable form of cosmic censorship obtains, then one can appeal to various "no hair" theorems for black holes to obtain a characterization of the final state of a gravitationally collapsed body in terms of a member of the two-parameter (mass, angular momentum) family of Kerr solutions. Second, the now standard black hole thermodynamics makes use of Hawking's area theorems, which in turn presuppose a form of cosmic censorship. And third, the proof of the positivity of the total mass of an isolated gravitating system (the Arnowitt–Deser–Misner (ADM) mass) presupposes the absence of singularities on the initial time slice; such an absence can be viewed as a kind of cosmic censorship. These matters will not be discussed here; the interested reader may consult standard texts on GTR such as Wald (1984a).

3.3 Naked singularities and cosmic censorship

When Penrose first raised the problem of cosmic censorship it was not clear what to include under the notion of a naked singularity. For example, Penrose said "In one sense, a 'cosmic censor' can be shown *not* to exist. For . . . the 'big bang' singularity is, in principle, observable" (Penrose 1969, p. 274). Today standard big bang cosmologies would *not* be regarded as nakedly singular and, thus, would *not* be regarded as being in conflict with cosmic

censorship. In what follows I will describe five attempts to pinpoint the correlative notions of cosmic censorship and naked singularities. The first approach seeks to supply necessary and sufficient conditions for cosmic censorship to hold; a naked singularity is then characterized indirectly as a condition that produces the violation of cosmic censorship. Other approaches proceed the other way round; they first attempt to give a direct definition of a naked singularity, and then subsequently define cosmic censorship in terms of the absence of naked singularities. The reader who dips into the scientific literature will find a seeming disrespect for the distinction between "What is cosmic censorship?" and "Is cosmic censorship true?" Formulations of cosmic censorship are often couched so as to maneuver around counterexamples. This is not necessarily a dishonest practice. Nor is it a practice unfamiliar to philosophers of science. Trying to provide a precise explication of a vague concept calls for decisions about how to draw the line among borderline cases. Since there are many plausible ways of scribing, it is not inappropriate for ulterior motivations—such as the desire to establish some form of cosmic censorship—to guide the pen. Cases of outright gerrymandering will be flagged.

Approach 1

Since a key concern with the development of naked singularities is the breakdown of predictability and determinism, cosmic censorship may be formulated by imposing conditions that assure that no such unwanted behavior occurs.

For future reference it is useful to begin with two definitions. A *time slice* of the spacetime M, g_{ab} is a spacelike hypersurface $\Sigma \subset M$ without edges. Such a Σ is the relativistic analogue of a Newtonian "constant time" slice. (Not every relativistic spacetime admits time slices; see chapter 6.) A time slice is said to be a *partial Cauchy surface* if it is achronal (i.e., not intersected more than once by any future-directed timelike curve).[4] A partial Cauchy surface is the appropriate platform on which to specify instantaneous initial data that, one may hope, will allow the future to be predicted and the past retrodicted. To help assure success in this regard, we can require that Σ is a *Cauchy surface* for M, g_{ab} in that the *total domain of dependence* $D(\Sigma) =: D^-(\Sigma) \cup D^+(\Sigma)$ of Σ is all of M. (This is equivalent to the definition given in section 2.6.) Alternatively, one could say that Σ is a *future* (respectively, *past*) *Cauchy surface* for M, g_{ab} just in case Σ partitions M and $D^+(\Sigma)$ (respectively, $D^-(\Sigma)$) contains all of that part of M that lies to the future (respectively, past) of Σ. Σ is a Cauchy surface just in case it is both past and future Cauchy.

The present approach leads to a strong and to a weak version of censorship. Strong cosmic censorship (SCC) holds for M, g_{ab} just in case M, g_{ab} possesses a Cauchy surface. Of course, it is assumed that M, g_{ab} is maximal, otherwise it would not represent a physically reasonable model (recall Fig. 2.5). The standard big bang cosmological models are maximal

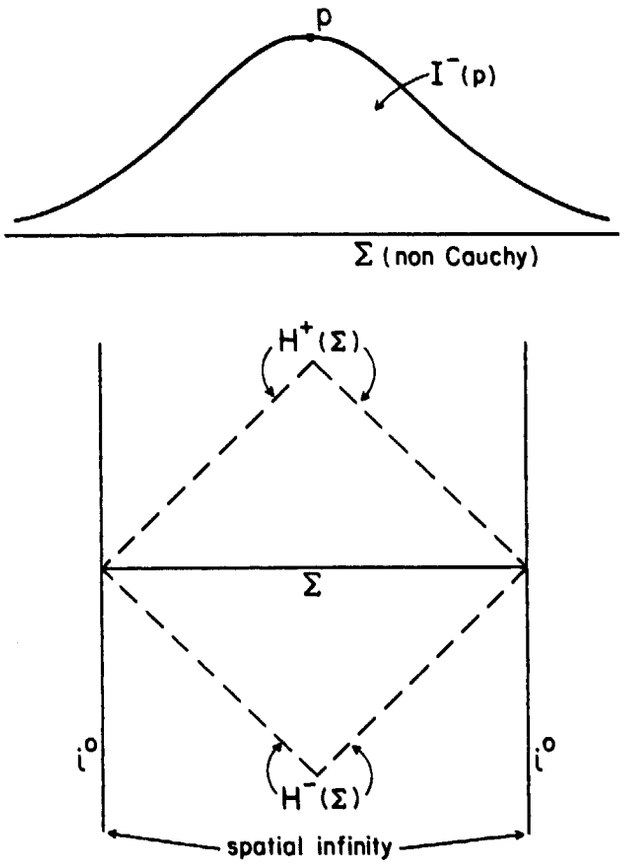

Fig. 3.3 (a) The behavior of light cones in anti-de Sitter spacetime; (b) Penrose conformal diagram of anti-de Sitter

globally hyperbolic so they satisfy the present statement of SCC, and thus the big bang singularity is *not* counted as naked.

Global hyperbolicity or the existence of a Cauchy surface is a very strong condition, and one can wonder whether it is too heavy-handed a way to capture the intuitive idea that there are no naked singularities. A relevant example is provided by the universal covering of the anti-de Sitter spacetime, a typical light cone of which is represented in Fig. 3.3a and whose Penrose diagram is given in Fig. 3.3b.[5] This spacetime violates the proposed formulation of SCC but it is arguably singularity free, e.g., it is geodesically complete. The proponent of the present approach could concede the point while maintaining that since the key worry raised by naked singularities is the breakdown in predictability and determinism, no harm is done by formulating a version of cosmic censorship that is strong enough to assuage the worry

whatever the source, naked singularities or no. Even if we share this sentiment, there is still an evident need for a separate definition of naked singularity. The second approach described below takes up this challenge.

We also want to be able to distinguish breakdowns in predictability that aren't too drastic. In particular, Nature may practice a form of modesty by hiding singularities in "black holes," the exterior regions of which may be predictable because they admit future Cauchy surfaces. If so, weak cosmic censorship (WCC) is said to hold. In asymptotically flat spacetimes this idea can be made precise by appealing to the notion of *future null infinity* $\mathscr{I}^+$, the terminus of outgoing null geodesics that escape to spatial infinity.[6] The interior B of a *black hole* is then defined to be the complement of the causal past $J^-(\mathscr{I}^+)$ of $\mathscr{I}^+$, i.e., the events that cannot be seen from future null infinity. The boundary E of B is the *absolute event horizon*; this is the modesty curtain that hides the goings-on in the interior of the black hole from external observers. These concepts are illustrated in Kruskal spacetime (the maximal analytic extension of Schwarzschild spacetime as shown in Fig. 3.4a) and its Penrose diagram (Fig. 3.4b).[7] Frank Tipler (1979) has proposed a more general definition of a black hole that does not assume asymptotic flatness and is supposed to apply to any stably causal spacetime.[8] Until this or some substitute definition is shown to be satisfactory, we cannot claim to have a general formulation of WCC in the cosmological setting.

The Kruskal–Schwarzschild spacetime is not only an example of a black hole and WCC but it also displays SCC since it possesses a Cauchy surface. The difference between the strong and weak versions of cosmic censorship is illustrated schematically in Fig. 3.5. In Fig. 3.5a the singularity that develops in gravitational collapse is hidden from external observers but is visible to observers within the black hole. In Fig. 3.5b the black hole is even blacker because even those unfortunate observers who fall into the hole cannot 'see' the singularity, though they may well feel and, indeed, may be torn apart by the tidal forces as they are inevitably sucked into the singularity. Nevertheless, they may take some cold comfort in the fact that SCC holds.

If SCC holds there is a sense in which, paradoxically, the singularity never occurs. It follows from a result of Geroch (1970b) that if M, g_{ab} admits a Cauchy surface then it also admits a *global time function*, a C^1 map $t: M \to \mathbb{R}$ which has a timelike gradient and which therefore increases along every future-directed timelike curve. And further, t can be chosen so that each $t =$ constant surface is Cauchy. Since no Cauchy surface can intersect the singularity, we can conclude that there is no time t at which the singularity exists. (Exercise for the reader: draw a foliation of Cauchy surfaces for the spacetime pictured in Fig. 3.5b.) Of course, the statement that no singularity exists will be hotly disputed by the ghosts of the observers who have been sucked into the black hole and have been snuffed out after only a finite existence (as measured by their respective proper times).

It is instructive to take the time reverses of the processes pictured in Fig. 3.5 to produce "white holes" where the singularities explode into expanding

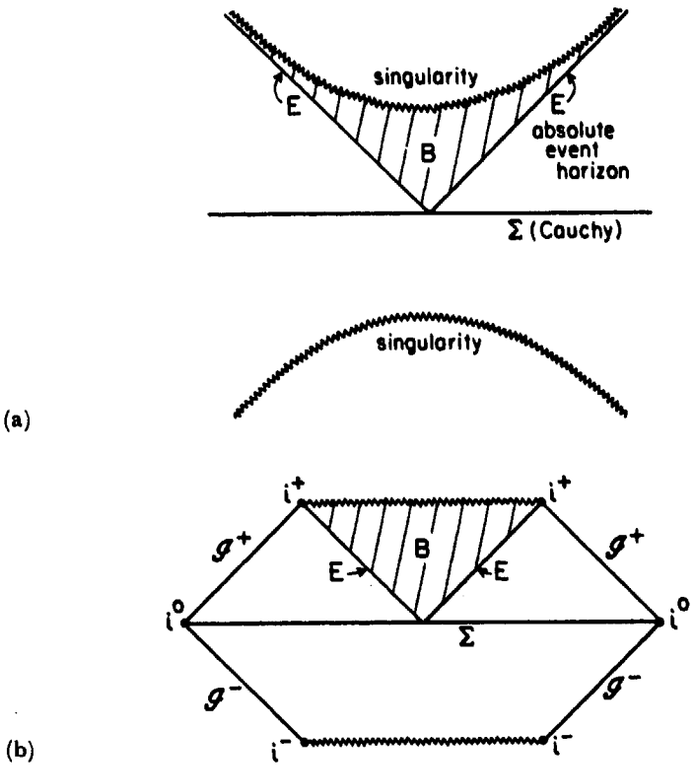

Fig. 3.4 (a) Kruskal–Schwarzschild spacetime; (b) conformal diagram of Kruskal–Schwarschild spacetime

ordinary matter (Fig. 3.6). At first glance the spacetimes of Fig. 3.6 would seem to possess naked singularities par excellence. But since the spacetime of Fig. 3.5b possess a Cauchy surface, so does the spacetime of Fig. 3.6b. In this sense the singularity in Fig. 3.6b is not naked even though it is highly visible (as is the case with the big bang singularity of the Friedmann–Robertson–Walker models).

But does predictability really hold in the situation pictured in Fig. 3.6b? Penrose has argued for a negative answer:

> The future behavior of such a white hole does not, in any *sensible* way, seem to be determined by its past. In particular, the precise moment at which the white hole explodes into ordinary matter seems to be entirely of its own 'choosing', being unpredicatble by the use of the normal laws of physics. (Penrose 1979, p. 601)

Penrose's point seems to be this. The spacetime in Fig. 3.6b can be foliated by Cauchy surfaces. But the singularity lies to the past of any such surface,

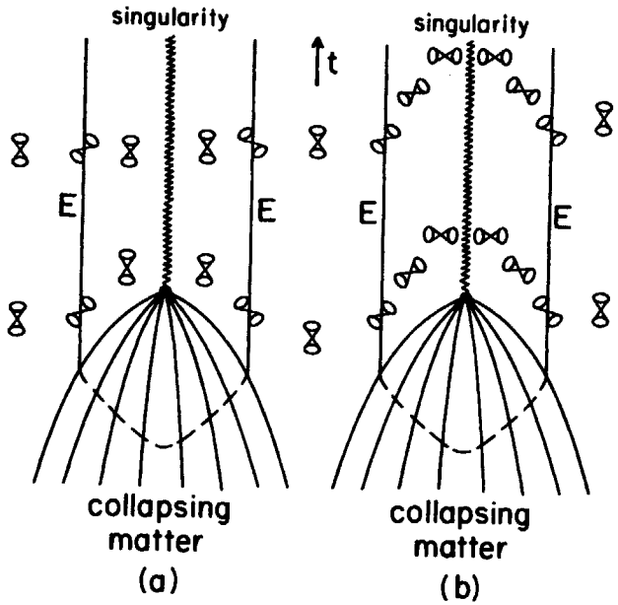

Fig. 3.5 Black holes

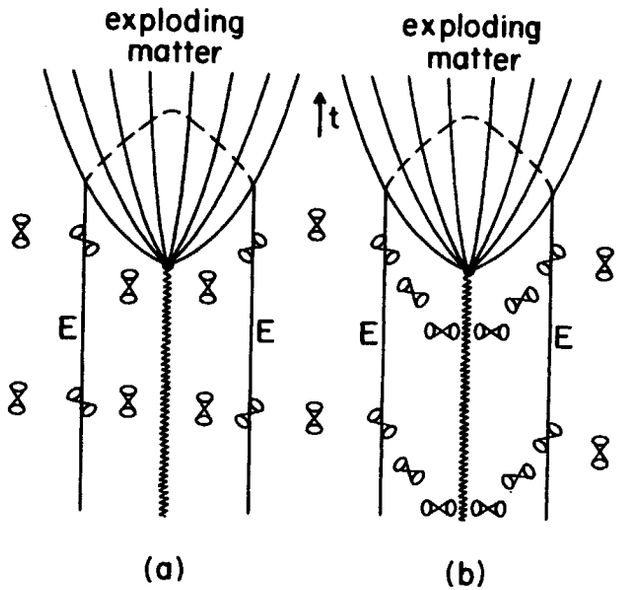

Fig. 3.6 White holes

which means that any such surface must intersect the ordinary matter. So the explosion cannot properly be said to be predicted from any such surface. There are other spacelike surfaces from which one can properly speak of predicting the explosion since it lies to their future. But since these surfaces do not have the Cauchy property, the prediction cannot be of the deterministic variety. If Penrose's worry is to be taken seriously and if it is to be assuaged by some form of cosmic censorship, then an even stronger version of SCC than the one given above is needed. This way seems to end in disaster for cosmic censorship, so I will not pursue the matter.

Figures 3.5 and 3.6 also raise another problem, not so much for the statement of cosmic censorship as for the validity and proof of the hypothesis that GTR contains a mechanism for enforcing cosmic censorship. Suppose that in typical cases of gravitational collapse SCC fails while WCC holds, i.e., the process in Fig. 3.5a is what we should expect. Then since Einstein's field equations are time reversal invariant, every solution of type 3.5a is matched by a solution of type 3.6a. So if black holes that violate SCC but satisfy WCC are a pervasive feature of general relativistic models, it would seem that white holes that violate WCC are also a pervasive feature.

One can take the attitude that what is needed here is a division of labor. The initial effort should be devoted to proving (or refuting) the conjecture that naked singularities do not occur in reasonable models of gravitational collapse. Then attention can be turned to the problem of white holes, which may be regarded as an aspect of the more general problem of time's asymmetries. In other branches of physics— electromagnetism and mechanics, for example—the fundamental laws are also time reversal invariant. But we find that while certain types of solutions are commonly encountered, their time-reversed counterparts never or very seldomly occur (e.g., we often encounter electromagnetic waves expanding from a center to spatial infinity but we never encounter waves converging on a center). So it is hardly unexpected that an analogous situation occurs in gravitational physics. The fact that the origin of time's arrow remains an unsolved problem is nothing to boast about, but it is not a special problem for gravitational physics and it should not prevent work from going ahead on the issue of cosmic censorship.[9]

Although the present approach to cosmic censorship has yielded some valuable insights, it is subject to some serious shortcomings. Most obviously, although it provides sufficient conditions (disregarding Penrose's worry) for ruling out naked singularities, it hasn't told us directly what a naked singularity is, nor has it told us how the violation of cosmic censorship leads to singularities in some intuitive sense.

Work by Joshi and Saraykar (1987) provides some information on the latter issue. We know from results discussed earlier that SCC in the guise of the existence of a Cauchy surface Σ implies that the topology of space does not change with time in the sense that the spacetime manifold is diffeomorphically $\Sigma \times \mathbb{R}$. So it is natural to ask: If SCC is violated and there is a

change in topology, what can we expect about the existence of singularities? To make this question more precise, define a partial Cauchy surface Σ of M, g_{ab} to be *maximal* just in case $D(\Sigma)$ is maximal in the set of $D(\Sigma')$ for all partial Cauchy surfaces Σ' for M, g_{ab}. If Σ and Σ' are both maximal, $\Sigma' \subset I^+(\Sigma)$, and $\Sigma' \not\subseteq \Sigma$ then a topology change is said to take place.[10] Since Σ cannot be a Cauchy surface, $H^+(\Sigma)$ cannot be empty. Joshi and Saraykar show that if in addition the weak energy condition is satisfied and if all timelike trajectories encounter some non-zero matter or energy, then a singularity will occur in that a null generator of $H^+(\Sigma)$ will be past incomplete.[11]

Nevertheless, it would be nice to have a direct definition of 'naked singularity' entailing that singularities cannot exist if there is a Cauchy surface. This would confirm that the approach explored above is on the right track.

Approach 2

The second approach seeks to define a naked singularity in terms of its detectability. Cosmic censorship then becomes the statement that such singularities do not occur.

Penrose's (1974, 1978, 1979) version of this approach emphasizes local detectability.

> It seems to me to be quite unreasonable to suppose that the physics in a comparatively local region of spacetime should really 'care' whether a light ray setting out from a singularity should ultimately escape to 'infinity' or not. To put things another way, some observer . . . might intercept the light ray and see the singularity as 'naked', though he be not actually situated at infinity. . . . The unpredictability entailed by the presence of naked singularities which is so abhorrent to many people would be present just as much for this local observer . . . as for an observer at infinity.
>
> It seems to me to be comparatively unimportant whether the observer himself can escape to infinity. Classical general relativity is a scale-independent theory, so if locally naked singularities can occur on a very tiny scale, they should also, in principle, occur on a very large scale. . . .
>
> It would seem, therefore, that if cosmic censorship is a principle of Nature, it should be formulated in such a way as to preclude such *locally* naked singularities. (Penrose 1979, pp. 618–619)

Penrose's technical explication of the notion of a locally naked singularity uses the concepts of TIFs and TIPs which I do not want to introduce here,[12] so I will follow a related idea used by Geroch and Horowitz (1979). The form of a definition for the set $\mathfrak{N}$ of points from which the spacetime M, g_{ab} can be detected to be singular can be stated as follows:

DEFINITION 3.1
$\mathfrak{N} = \{p \in M: I^-(p)$ contains a timelike curve γ which has no future endpoint and which _____$\}$

The reason, intuitively speaking, that γ has no future endpoint is that it runs into a singularity. Since this fact is directly detectable from p, the spacetime is nakedly singular as viewed from p. Thus, the statement that M, g_{ab} harbors naked singularities becomes: $\mathfrak{N} \neq \varnothing$; and conversely, cosmic censorship becomes the statement: $\mathfrak{N} = \varnothing$.[13]

The strongest version of cosmic censorship is obtained by putting no further restrictions in the blank. Then $\mathfrak{N} = \varnothing$ is equivalent to the existence of a Cauchy surface, a result that dovetails nicely with Approach 1. If we are somewhat more restrictive and fill the blank with "is a geodesic," then $\mathfrak{N} = \varnothing$ no longer entails the existence of a Cauchy surface. Depending upon one's point of view, it could be counted as a benefit of this latter version of cosmic censorship that anti-de Sitter spacetime is no longer counted as nakedly singular. (Referring back to the rendering of the null cone behavior in anti-de Sitter spacetime in Fig. 3.3a, consider an arbitrary spacetime point p. A timelike geodesic that starts at a point $q \in I^-(p)$ will eventually escape $I^-(p)$ if it is extended far enough into the future.)

If one wants the focus of cosmic censorship to be curvature singularities, then the blank should be filled with restrictions that guarantee γ terminates on what one chooses to regard as a curvature singularity. Other fillings would be appropriate depending upon what kinds of naked singularities one wants to target. In this way we obtain several versions of cosmic censorship.

If one does not share Penrose's sentiments about local detectability, the present approach can still be adapted to a weaker statement of cosmic censorship by requiring only that $\mathfrak{N} = \varnothing$ for the region exterior to black holes.

Approach 3

As part of establishing cosmic censorship, one would like to prove that in reasonable models of gravitational collapse naked singularities do not develop from regular initial data. Thus, on this approach one would not regard the negative mass Schwarzschild model[14] as a counterexample to cosmic censorship since, although it is nakedly singular, the singularity has been around for all time.

Before we can start on this project we need a definition to isolate naked singularities that can be said to develop from regular initial data, as illustrated schematically in Fig. 3.7. The ideas of Geroch and Horowitz (1979) and Horowitz (1979) suggest:

DEFINITION 3.2
A spacetime M, g_{ab} is *future nakedly singular in the first sense* (FNS_1) with respect to the partial Cauchy surface $\Sigma \subset M$ just in case there is a $p \in H^+(\Sigma)$ such that $\overline{I^-(p)} \cap \Sigma$ is compact.

Newman (1984b) works with a somewhat stronger criterion:

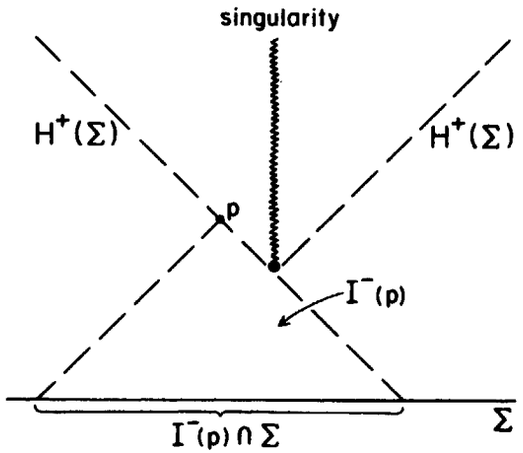

Fig. 3.7 The formation of a naked singularity from regular initial data

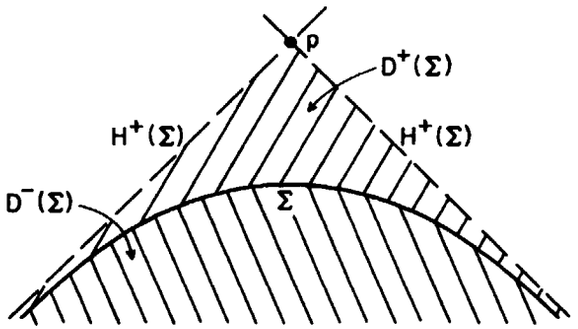

Fig. 3.8 A poor choice of initial value hypersurface

DEFINITION 3.3
A spacetime M, g_{ab} is *future nakedly singular in the second sense* (FNS_2) with respect to the partial Cauchy surface $\Sigma \subset M$ just in case there is a $p \in H^+(\Sigma)$ such that $\overline{I^-(p)} \cap \Sigma$ is compact and a null geodesic generator of $H^+(\Sigma)$ through p is past incomplete.

The implicit assumption of this approach is that all of the spacetimes under discussion are maximally extended; for if M, g_{ab} were, say, the maximal Cauchy development of initial data on Σ, then it would automatically be globally hyperbolic and $H^+(\Sigma)$ would be empty. The condition that $\overline{I^-(p)} \cap \Sigma$ is compact assures that Minkowski spacetime is not counted as being FNS_1 or FNS_2 with respect to the Σ illustrated in Fig. 3.8 (compare to Fig. 3.7). However, as will be seen below, this condition has some other not-so-nice consequences. Anti-de Sitter spacetime (Fig. 3.3) is

not future nakedly singular in either sense FNS_1 or FNS_2. The two-dimensional Misner spacetime of Fig. 2.3, illustrating some of the causal features of Taub–NUT spacetime, is obviously FNS_1 with respect to Σ. (The surface labeled $\mathcal{N}$ is in fact $H^+(\Sigma)$. For any $p \in I^-(\Sigma)$, $\Sigma \subset I^-(p)$; and since Σ is compact, $I^-(\Sigma) \cap \Sigma$ is compact.) It is not FNS_2; indeed, if $H^+(\Sigma)$ is compact for a partial Cauchy surface Σ, then the null geodesic generators of Σ are past complete; see Hawking and Ellis (1973, Lemma 8.5.5).

Definition 3.2 is more appropriate for capturing the notion of a breakdown in predictability or determinism due to whatever cause, whereas Def. 3.3 is more relevant to identifying the development of naked singularities of the type that are indicated by geodesic incompleteness. But as noted in chapter 2, geodesic incompleteness is not a reliable indicator of a curvature singularity. So if it is curvature singularities that are to be ruled out by the statement of cosmic censorship, a further definition is needed which strengthens Def. 3.3 by requiring that, in some appropriate sense, a generator of $H^+(\Sigma)$ through p encounters a curvature singularity of some specified type (e.g., scalar polynomial blowup) in the past direction.

Before proceeding further one should make sure that a future nakedly singular spacetime as defined above really is singular in some minimal sense. This would mean showing that if M, g_{ab} is future nakedly singular with respect to Σ, then the region of M that is on the future side of Σ and that is causally accessible from Σ is not globally hyperbolic. Suppose for purposes of contradiction that $J^+(\Sigma)$ is globally hyperbolic. Then on either Definition 3.2 or 3.3 there is a point $p \in H^+(\Sigma)$ such that $\overline{I^-(p)} \cap \Sigma$ is compact. So $J^-(p) \cap J^+(\overline{I^-(p)} \cap \Sigma)$ is a compact set (Hawking and Ellis 1973, Cor. to Prop. 6.6.1). Since Σ has no edge, the generator α of $H^+(\Sigma)$ through p is past endless. And since $\alpha \in (J^-(p) \cap J^+(\overline{I^-(p)} \cap \Sigma))$ we have a past endless null curve imprisoned in a compact set. Thus $J^+(\Sigma)$ is not strongly causal and, a fortiori, is not globally hyperbolic (Hawking and Ellis 1973, Prop. 6.4.7).[15]

Conversely, can we be sure that if M, g_{ab} is not future nakedly singular with respect to Σ then $J^+(\Sigma)$ is globally hyperbolic? The answer is yes if Σ is compact and "future nakedly singular" is taken as FNS_1. Indeed, Σ itself must be a Cauchy surface. If Σ were not a Cauchy surface $H^+(\Sigma)$ would be non-empty. Since Σ is compact, $\overline{I^-(p)} \cap \Sigma$ is compact for any $p \in H^+(\Sigma)$, so that the spacetime is FNS_1 with respect to Σ. On the other hand, this result does not hold if Σ is not compact, as we already know from the case of anti-de Sitter spacetime. But an even worse counterexample is provided by Reissner–Nordström spacetime (a piece of whose Penrose diagram is shown in Fig. 3.9) since this model possesses a naked curvature singularity to the future of a partial Cauchy surface. One could argue, however, that the very feature which allows this spacetime to count as neither FNS_1 nor FNS_2 makes it physically unreasonable (see Wald 1984a, p. 304). Namely, for any partial Cauchy surface Σ lying below the singularity and for any $p \in H^+(\Sigma)$, $I^-(p)$ contains an infinite (i.e., non-compact) portion of Σ. As a result a small perturbation on Σ can accumulate on $H^+(\Sigma)$ to produce an "infinite blue-shift"

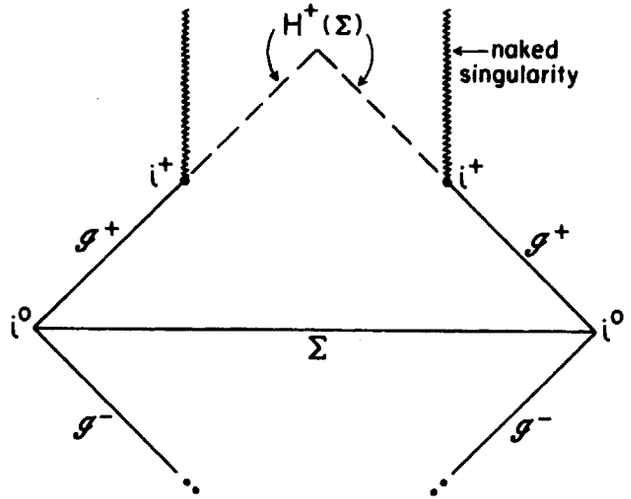

Fig. 3.9 A conformal diagram of part of Reissner–Nordström spacetime

singularity (Chandrasekar and Hartle 1982). This line of reasoning, however, has the defect of failing to respect a division of labor between framing a definition of what it is for a spacetime to be future nakedly singular and distinguishing between physically reasonable versus physically unreasonable violations of cosmic censorship. I warned in the outset that the division can become blurry in borderline cases; but the present case seems to lie far from the border.

If we do want to try to respect the division of labor, we are faced with a version of the by now familiar tension. On the one hand we can try to target various forms of singular behavior that in some sufficiently tight sense are traceable to the development of initial data from Σ; that is the course taken in the above definitions. On the other hand we can try to frame the definition of 'future nakedly singular with respect to Σ' so that being *not* future nakedly singular with respect to Σ entails that $J^+(\Sigma)$ is free of any taint of singularity. But the only way to assure the latter is to require that $J^+(\Sigma)$ is globally hyperbolic, and in turn the only sure way to guarantee that the development of initial data from Σ makes $J^+(\Sigma)$ globally hyperbolic is to require that $D^+(\Sigma)$ includes $J^+(\Sigma)$, which is to say that Σ is a future Cauchy surface and which is to collapse back to Approach 1.

Approach 4

Approach 3 was supposed to focus on how naked singularities might develop from regular initial data. But in fact it ignored the details of the initial value problem in GTR and implicitly assumed that the spacetimes at issue were maximal without inquiring whether they got that way by evolution of initial data or not. To remedy this defect, at least for an important subclass of cases,

consider the initial value problem for Einstein's vacuum field equations. A data set consists of a three-manifold Σ without boundary, to be realized as a spacelike hypersurface in the evolved spacetime; the *first fundamental form* h_{ab} of Σ, a Riemann metric that characterizes the intrinsic spatial geometry of Σ; and the *second fundamental form* K_{ab} which is (roughly) the time derivative of h_{ab}. To be consistent with EFE this data must satisfy a set of four coupled partial differential equations, called the *constraint equations*.[16] For any such data set, there exists a unique (up to isometry) spacetime M, g_{ab} which is a development of the initial data set,[17] which has Σ as a Cauchy surface, and which is inextendible in any way that keeps Σ a Cauchy surface (Choquet-Bruhat and Geroch 1969).

One might be tempted to say that if this M, g_{ab} is not maximal simpliciter then cosmic censorship is violated because an observer crossing over $H^+(\Sigma)$ into the region $\tilde{M} - M$ of an extension $\tilde{M}, \tilde{g}_{ab}$ of M, g_{ab} will enter *terra incognita* where the spacetime geometry is not determined by the initial data on Σ. However, the non-maximality of M, g_{ab} may not signal a lapse on the part of the cosmic censor but only a poor choice of the initial value hypersurface. For example, Σ might be chosen to be an open spacelike disk of Minkowski spacetime. This example can be ruled out by requiring that Σ be "large enough"; in particular, Σ, h_{ab} should be a complete Riemann space (see chapter 2). But the requirement of completeness is not strong enough to guarantee that Σ is a good choice in the appropriate sense. Consider the spacelike hyperboloid Σ of Minkowski spacetime as illustrated in Fig. 3.8. With the space metric h_{ab} induced by the Minkowski metric η_{ab}, the Riemann space Σ, h_{ab} is complete, but the maximal development for which Σ is a Cauchy surface is extendible. This example can be ruled out by the further requirement that when Σ is non-compact, Σ, h_{ab} should be asymptotically flat. But this further requirement is too strong for a general formulation of cosmic censorship since the censor may have to work in cosmologies that are neither spatially closed nor asymptotically flat.

These difficulties do not arise in the spatially closed case. For if Σ is compact, then Σ, h_{ab} must be complete. And, as already noted above, in a globally hyperbolic spacetime admitting a compact Cauchy surface, any compact time slice is a Cauchy surface (see Budic et al. 1978).

Approach 5

It is now time to question the obsession with Cauchy surfaces. To make the point it suffices to continue with the special case of vacuum solutions. Let Σ, h_{ab}, K_{ab} be an initial data set. And to sidestep the difficulties encountered in Approach 4 we may further specialize to the case where Σ is compact or else Σ, h_{ab} is complete and asymptotically flat. If the worry over cosmic censorship is that deterministic evolution may break down, then the worry is assuaged if it can be shown that there is a unique (up to diffeomorphism) maximal spacetime M, g_{ab} which is a solution of the vacuum field equations

and which is a development of the given initial data. And that is so regardless of whether M, g_{ab} contains a Cauchy surface or not. Of course, the expectation is that in typical cases when the unique maximal development for which Σ is a Cauchy surface is not maximal, then extensions across $H^+(\Sigma)$ will not be unique. But that expectation needs to be substantiated, otherwise formulations of cosmic censorship in terms of global hyperbolicity will loose their interest (see Chruściel 1992). I will temporarily set aside this concern, but will return to it in section 3.7.

3.4 The cosmic censorship hypothesis

The cosmic censorship hypothesis (CCH) is the claim that the only naked singularities that occur in the models of GTR are *harmless*.[18] Recall that a *model* of GTR is a triple M, g_{ab}, T^{ab}, where M, g_{ab} is a relativistic spacetime, T^{ab} is a symmetric second-rank tensor called the *stress–energy–momentum tensor*, and g_{ab} and T^{ab} together satisfy EFE. Examples of naked singularities in such models are easily constructed, as we know from the above discussion. But for one reason or another such examples may be brushed aside as harmless. The principal reason for putting a singularity in the harmless category is that the model in which it occurs has features that, apart from the singularity itself, make it "physically unreasonable." The literature on cosmic censorship can be confusing to the casual reader because it mixes together a number of different senses in which a model can be physically unreasonable, among them: the model is literally physically impossible; the model involves unrealistic idealizations, and there is no reason to expect that more realistic counterparts will also have naked singularities; the model is physically possible but involves such rare features as to leave no reason to think that anything like the model will be actually encountered.

Whatever specific content is given to the physically reasonable/physically unreasonable distinction, there are two boundary conditions that should be satisfied. First, 'physically unreasonable' should not be used as an elastic label that can be stretched to include any ad hoc way of discrediting putative counterexamples to the CCH. Second, the conception we settle on must permit the CCH to be stated in a precise enough form that it lends itself to proof (or refutation). Some aspects of the physically reasonable/unreasonable distinction can be stated in advance. Others emerge only in the process of assessing potential counterexamples to the CCH. This, of course, raises worry about the first boundary condition. But as we shall see below, the real worry is about satisfying the second boundary condition while at the same time making the CCH not obviously false and also general enough to cover the situations that can be expected to occur in the actual universe.

Among the constraints on a physically reasonable model of GTR there are two which, at least in part, can be motivated independently of any concern with cosmic censorship.

Energy conditions

Start with a spacetime M, g_{ab} and compute the Einstein tensor G_{ab} associated with g_{ab}. Then define $T_{ab} =: (1/8\pi)G_{ab}$. The result is a model that satisfies EFE with cosmological constant $\Lambda = 0$. In this way we create, at least formally, innumerable models of GTR, among which will be many that violate cosmic censorship.[19] To return to an example from section 3.1, we can start with empty Minkowski spacetime $\mathbb{R}^4, \eta_{ab}$, remove a compact set $K \subset \mathbb{R}^4$, and choose a conformal factor Ω that goes rapidly to 0 as K is approached.[20] The resulting triple $M = \mathbb{R}^4 - K$, $g_{ab} = \Omega^2\eta_{ab}$, $T_{ab} = (1/8\pi)G_{ab}(g_{ab})$ is a nakedly singular model in which the singularity cannot be removed by extensions since the spacetime is maximal. Of course, we may not get a model in the intended sense that the stress–energy–momentum tensor arises from normal sources, such as massive particles, electromagnetic fields, and the like. This intention is difficult to state in terms of precise formal conditions on T^{ab}, but at a minimum we can impose one or another energy condition that we expect normal sources to satisfy.

In chapter 2 we encountered the weak and strong energy conditions which require respectively that $T_{ab}\zeta^a\zeta^b \geq 0$ for any timelike ζ^a and that $T_{ab}\zeta^a\zeta^b \geq -(1/2)T$, $T =: T^a{}_a$, for any unit timelike ζ^a. The *dominant energy condition* requires that for any timelike ζ^a, $T^a{}_b\zeta^b$ is a future-directed timelike or null vector, which means that the flow of energy–momentum as measured by any observer does not exceed the speed of light.[21] To see what these conditions mean for a concrete example, consider a perfect fluid whose stress–energy–momentum tensor has the form $T_{ab} = (\mu + p)U_aU_b + pg_{ab}$, where μ and p are respectively the energy density and pressure of the fluid and U^a is the unit tangent to the world lines of the fluid elements. The strong energy condition says that $\mu + p \geq 0$ and $\mu + 3p \geq 0$. The weak energy condition requires that $\mu \geq 0$ and $\mu + p \geq 0$. And the dominant energy condition says that $\mu \geq 0$ and $\mu \geq p$.

I conjecture that these energy conditions, which are thought to hold for all physically reasonable classical fields, rule out all the artificial examples of naked singularities constructed by the method of two paragraphs above, but I know of no formal proof of this.

One can now appreciate the importance of the value assigned to the cosmological constant. Anti-de Sitter spacetime has constant scalar curvature $R < 0$ and Einstein tensor $G_{ab} = -(1/4)Rg_{ab}$. With $\Lambda = 0$ we can interpret this as a solution with a perfect fluid source of constant density $(-R/32\pi) > 0$ and constant pressure $(R/32\pi) < 0$. It is ruled out by the strong energy condition. If, however, the cosmological constant is allowed to be non-zero, then anti-de Sitter spacetime can be interpreted as an empty-space solution ($T_{ab} = 0$) with $\Lambda = (1/4)R$. Since this spacetime violates the strongest version of cosmic censorship, it might seem that without the stipulation of a zero cosmological constant it will be much more difficult to achieve strong cosmic censorship. But recall from chapter 1 that a positive value for Λ can help to

prevent the occurrence of singularities. So it remains to be seen whether or not, on balance, the cosmological constant helps or hinders the quest for cosmic censorship.

Causality conditions

For both physical and philosophical reasons one might require that a reasonable model not contain closed or almost closed causal curves.[22] But imposing causality conditions by fiat is a little awkward for present purposes since it smacks of assuming what we want to prove, at least for versions of cosmic censorship that seek to censure breakdowns of predictability and determinism that occur because of the development of acausal features after the specification of initial data. Thus, in Taub–NUT spacetime we can choose a partial Cauchy surface Σ lying in the Taub region such that in some neighborhood of Σ things are causally as nice as we like; but further to the future of Σ things turn causally nasty, and as a result a Cauchy horizon for Σ develops (recall Fig. 2.3). At the present stage of discussion, however, fiat is needed to secure cosmic censorship since Taub–NUT spacetime is a vacuum solution to EFE so that the energy conditions are trivially satisfied. (In chapter 6 I will take up the question of whether or not GTR allows the operation of time machines which would manufacture closed timelike curves. A negative answer would establish part of the cosmic censorship hypothesis. Conversely, a positive answer would be very damaging to the hypothesis.)

With only energy and causality conditions in place, the CCH fails, as is shown by the "shell-crossing" singularities that may arise in spherically symmetric dust collapse (see Yodzis et al. 1973). The collapse is arranged so that the outer shells of dust fall inward faster than the inner shells. A black hole eventually develops, but not before the crossing of shells produces an infinite density singularity that is visible from both near and far. Hawking (1979) and others have suggested that such naked singularities are harmless because they are relatively mild; in fact, the solution can be continued through the Cauchy horizon as a generalized distributional solution (Papapetrou and Hamoui 1967). However, Seifert (1979) has cautioned that harmlessness in the relevant sense has not been demonstrated if our concern with cosmic censorship is over the potential breakdown in predictability and determinism, for uniqueness theorems for such generalized solutions are not in place. In any case this method of trying to render naked singularities harmless does not have a very long reach since, as will be discussed below, stronger and irremovable singularities threaten.

Further conditions on T^{ab}

A second strategy for dealing with shell focusing singularities and similar examples is to impose further conditions on T^{ab} and the equations of state. These further conditions are supposed to assure that the sources are sufficiently

realistic. In particular, it could be demanded that pressure effects (neglected in dust models) be taken into account, and further that if matter is treated as a perfect fluid, the equation of state must specify that $p = p(\mu)$ is an increasing function of density and that the pressure becomes unbounded as the density becomes unbounded. This would rule out some of the early shell-focusing counterexamples. But energy, causality, and further conditions are not sufficient to prevent violations of cosmic censorship in self-similar gravitational collapse[23] with an equation of state $p = a\mu$, with $0 < a < 1$ constant (see Ori and Piran 1987, 1988, 1990). Perhaps reasons can be found to deem this soft equation of state physically unreasonable. Or perhaps it should be demanded that to be realistic the fluid description should incorporate viscosity and, thus, shear stress into T^{ab} (see Seifert 1983). Further violations of cosmic censorship satisfying these additional demands would drain much of the interest in this tack for trying to render potential counterexamples harmless.

Eardley's conditions

A related but different tack is taken by Eardley (1987), who demands realistic equations of motion for the sources. In analyzing the naked shell-focusing singularities that emerge in the Tolman–Bondi models of spherical gravitational collapse as matter piles up at the center of symmetry, Eardley found that the tacit assumption underlying the model is that the dust shells cannot pass through the origin. The objection is not to the idealization of matter as a pressureless, viscosity-free dust but rather to the unreasonable assumption that dust shells behave completely inelastically. He conjectures that if the motion of the dust is treated more realistically, for example by specifying elastic recoil when the shells collide at the origin, then naked singularities will not develop. Settling this and related conjectures seems to me to be one of the more important items on the agenda for evaluating the prospects of the CCH. However, I worry that the present approach, while possibly effective in dealing with putative counterexamples to cosmic censorship on a case-by-case basis, may not lead to a neat formal statement of the CCH that can be proved to hold for a wide class of models.

Fundamental fields

Yet another approach in the spirit of Further conditions and Eardley's conditions would recognize counterexamples to cosmic censorship only in models involving fundamental fields. The gravitational field itself is, of course, a fundamental field. The electromagnetic field should also be counted as fundamental. And presumably so should a scalar field obeying a Klein–Gordon type equation. Two aspects of determinism unite these examples. First, the combined gravitational–matter fields for these cases admit of a locally well-posed initial value problem. (This is the motivation behind the

part of Wald's (1984a, p. 305) formulation of cosmic censorship that requires that the coupled Einstein–matter equations can be put in the form of a second-order, quasi-linear, diagonal hyperbolic system, for which local existence and uniqueness theorems are available.) Second, the non-gravitational field should admit a globally well-posed initial value problem in Minkowski spacetime. The imposition of fundamental fields would disqualify many potential counterexamples to cosmic censorship, including those involving perfect fluids.

The motivation for the restriction of fundamental fields on physically reasonable models lies in the notion that if either of the above determinism requirements fails, then the field description involves a false idealization or otherwise fails to be detailed and precise enough. The underlying faith here is that in classical GTR it is always possible to go down to a fundamental enough level of description on which the determinism requirements can be satisfied.[24] Of course, this article of faith could turn out to be wrong. But it is polemically buttressed by the fact that the question of cosmic censorship would be much less interesting if determinism could fail in special and general relativistic settings for reasons having nothing to do with the development of naked singularities.

Strength of singularities

Tipler (1985) and Newman (1986) have suggested a way to avoid the delicate and contentious question of what counts as a physically reasonable source by concentrating instead on strength of singularities. The idea is that for a source to create a physically realistic singularity, the singularity must be strong enough to crush ordinary matter out of existence by squashing it to zero volume. A formal definition of the shrinking to zero volume was provided by Tipler (1977b). Clarke and Krolak (1985) showed that a sufficient condition for such behavior to occur along an incomplete timelike or null geodesic is that $\lim_{\lambda \to \lambda^+} \lambda^2 R_{ab} V^a(\lambda) V^b(\lambda) \neq 0$ where $V^a(\lambda)$ is the tangent to the geodesic, λ is an affine parameter, and λ^+ is the upper bound on λ. The term *strong curvature singularity* is sometimes used to denote a singularity such that at least one timelike or null geodesic terminating in the singularity satisfies the crushing condition. Tipler (1977b) originally reserved the term for cases where all non-spacelike geodesics terminating in the singularity satisfy the crushing condition. Presumably in the latter cases all strong curvature singularities are scalar polynomial singularities (see section 2.4); but I know of no proof to this effect.

On the present approach we can give the CCH a clean and precise formulation: the *strong* (respectively, *weak*) *CCH* holds just in case strong (weak) censorship holds in any model which satisfies the energy conditions and the causality conditions and in which the only singularities that occur are of the strong curvature type. Unfortunately, the virtues of simplicity and precision are not rewarded by truth, for this version of the CCH is in fact

false, as is shown by the presence of strong naked singularities in the models of self-similar gravitational collapse (Lake 1988; Ori and Piran 1990).

Penrose's stability constraint

The final constraint on physically reasonable models I will consider is Penrose's (1974, 1978) idea of stability under perturbations of initial conditions and equations of state. The idea is often illustrated by an example mentioned in the preceding section, the Reissner–Nordström model, where the effects of small perturbations on a partial Cauchy surface Σ in the initial data for the Einstein–Maxwell equations can build until they become an infinite blueshift singularity on $H^+(\Sigma)$. But exactly what does such a demonstration show? Wald takes it to show "there is good reason to believe that in a physically reasonable case where the shell is not exactly spherical, the Cauchy horizon ... will become a true, physical singularity, thereby producing an 'all encompassing' singularity inside the black hole formed by the collapse" (Wald 1984a, p. 318). But if this is the moral of the perturbational analysis then one should be able to drop reference to instability under perturbations and say directly (either in terms of the above ideas or in terms of some altogether different ideas) what a physically reasonable model is and then proceed to prove that physically reasonable models so characterized obey cosmic censorship.

I take it, however, that what the talk about instability is supposed to point to is the notion that naked singularities that develop from regular initial data are relatively rare within the set of all models of GTR. To make this precise one would need to define a topology on the set of solutions to Einstein's field equations and show that the target set is the complement of an open dense set. Given our present limited knowledge of generic features of solutions to Einstein's field equations, the project of establishing this version of the CCH seems rather grandiose. In the meantime we can lower our sights and investigate particular families of models whose members can be parameterized in a natural way and try to show that cosmic censorship holds for almost all parameter values. Or failing this we can challenge potential counterexamples by showing instability under small perturbations and take this as a sign that in some sense yet to be made precise a measure-zero result should be forthcoming. The bad news here is that various counterexamples have been shown to be free of blueshift instabilities, and the formation of naked singularities in some cases of spherical collapse have been shown to be stable under spherically symmetric perturbations. If it can also be shown that the formation of naked singularities is stable under generic perturbations, then we will have a strong indication that solutions with naked singularities are not rare beasts, and the CCH will have to avail itself of some other escape route. If on the contrary it can be shown that solutions with naked singularities are rare beasts, there still remains the nagging question of what we would say if we found that the actual universe we inhabit is one of the rare ones.

The good news for philosophers of science from this review is that the hunting ground for the CCH contains a rich array of examples for studying the related concepts of 'physically possible', 'physically reasonable', and 'physically realistic'. The bad news for advocates of cosmic censorship is that the CCH does not yield easily to a formulation that is not obviously false, is reasonably precise, and is such that one could hope to demonstrate its truth. There is the very real danger that the CCH will forever remain a hypothesis, and a rather vague hypothesis at that.

3.5 Is the cosmic censorship hypothesis true?

Since the literature does not agree on a precise statement of the CCH, the question is a murky one. Not surprisingly, the murk has provided fertile soil from which have sprung a wide variety of clashing opinions. Rather than review these opinions, it seems to me more productive to try to state the strongest cases on either side of the issue.

On the negative side we can begin by noting that faith in the CCH depends on the notion that the GTR has some built-in mechanism for preserving modesty by clothing singularities. In this respect the original Oppenheimer–Snyder–Volkoff model of gravitational collapse (Oppenheimer and Snyder 1939; Oppenheimer and Volkoff 1939) was misleading in creating a false sense of security. Indeed, subsequent analysis has revealed a number of different mechanisms by means of which cosmic censorship can be violated. Steinmuller, King, and Lasota (1975) showed that a momentarily naked singularity (see Fig. 3.10) can be produced by having a collapsing star radiate away its mass in such a way that the star never forms a black hole since it remains outside of its Schwarzschild radius (see also Lake and Hellaby 1981).

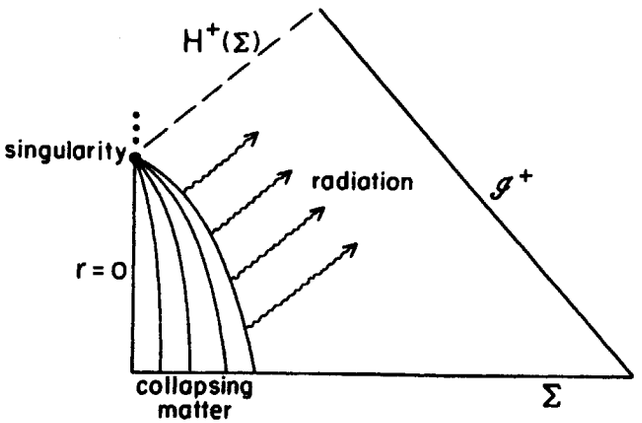

Fig. 3.10 A naked singularity emerges from a collapsing, radiating star

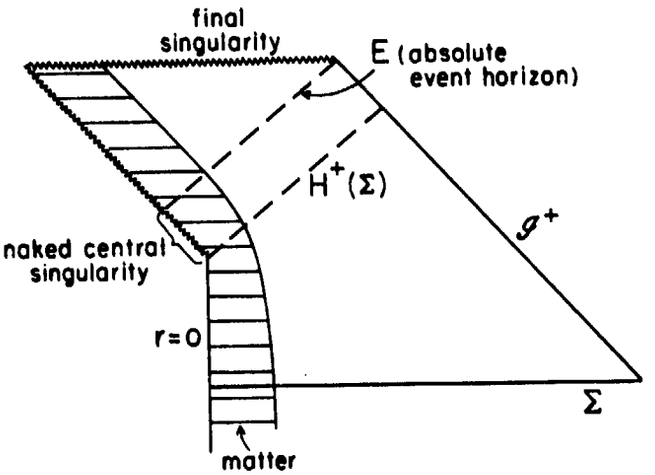

Fig. 3.11 Shell-focusing in self-similar gravitational collapse leads to a naked singularity

Then there are the persistently naked strong curvature singularities that form in self-similar gravitational collapse, either as a result of the gravitational collapse of a spherical shell of radiation as in the Vaidya spacetime, or as a result of the shell-focusing of dust in the Tolman–Bondi spacetimes (see Fig. 3.11), or as a result of the collapse of an adiabatic perfect fluid (see Joshi 1993, Ch. 6 for a review). Joshi and Dwivedi (1992) showed that non-self-similar gravitational collapse of a shell of radiation can also form strong curvature singularities that violate WCC. And Joshi (1993, Ch. 6) found that shell-focusing singularities of dust in non-self-similar gravitational collapse can also be strong and naked. To be sure, each of these examples involves special features and idealizations. But Joshi (1993, Ch. 7) makes it plausible that for self-similar collapse, the development of strong naked singularities is not due to the idealizations of perfect fluids, pressureless dust, and pure radiation but holds for any form of matter satisfying the weak energy condition. In any case, the fact that naked singularities can arise in such a variety of ways should shake one's faith that GTR does have a modesty mechanism at work.

This faith suffers another apparent blow from the computer simulation studies of Shapiro and Teukolsky (1991a). They interpret their results to indicate that a naked singularity can emerge from the collapse of a prolate spheroid of a collisionless gas.[25] However, this interpretation is currently a matter of dispute. In particular, Shapiro and Teukolsky take the absence of an apparent horizon in their simulations to serve as evidence for the absence of an event horizon clothing the singularity.[26] Wald and Iyer (1991) showed that such an inference is dangerous since in Schwarzschild–Kruskal spacetime there is a family of time slices that passes as near to the black hole singularity as you like and for which there is no apparent horizon although, of course,

event horizon. Another reason for doubting the significance of the Teukolsky result comes from their use of the Vlasov equation to distribution of dust particles in their model. Rendall (1992b) it is unreasonable to expect that more general solutions for this will exhibit the singular behavior found in the Shapiro–Teukolsky that its generality will in any case the use of computer models provides a powerful tool exploration of the cosmic censorship hypothesis, and given the difficulty progress in discovering counterexamples to various versions of censorship will be achieved with the help of this tool.)

Penrose (1973) noted that the CCH can be used to derive involving areas of trapped surfaces and available masses, on the us turn the coin to look at the evidence in favor of cosmic of black holes.[27] He tried but failed to find physically plausible ways these inequalities. The failure constitutes only weak evidence for censorship since, as Penrose himself notes, if cosmic censorship does a naked singularity may form without a trapped surface also Waald's (1973) investigation starts with the inequality necessary for Newman solutions to represent a black hole: $M^2 \geq Q^2 + \mathcal{J}^2/M^2$, Q, and $\mathcal{J}$ are respectively the mass, electric charge, and the angular He found that various ways of trying to violate this inequality charge and angular momentum into the black hole all fail. Again only weak support to the CCH since although it confirms the of stationary black holes once they form, it gives no confidence that rather than a naked singularity will form in gravitational collapse. (1986, 1987a, 1987b) has offered various censorship theorems, rely on the dubious assumption of the existence of marginally null geodesics.[28]

Newman (1984b) established a censorship theorem for conformally flat $(M, g_{ab}$ is *conformally flat* just in case $g_{ab} = \Omega^2 \eta_{ab}$ where η_{ab} has Riemann curvature and Ω^2 is a C^∞ map from M to $(0, +\infty)$.) In null convergent ($R_{ab}K^aK^b \geq 0$ for any null vector K^a) conformally all null geodesics are incomplete. Nevertheless, Newman if Ω^2 is a proper map (i.e., the inverse image of any compact is a compact subset of M) then M, g_{ab} is not FNS_2. Of more another result of Newman (1984a, 1986) that is another form of cosmic censorship for weakly asymptotically simple which satisfy a suitable causality condition and in incomplete null geodesic experiences a persistent curvature of However, the applicability of the persistent curvature and cosmology remains uncertain.

Isenberg, and Moncrief (1990) have established a form of gravitational collapse and cosmology for a special class of Gowdy spacetimes. These are censorship for a special class of Gowdy spacetimes. These are solutions to Einstein's field equations with vanishing cosmological They are spatially closed with space sections Σ being topologically

T^3, S^3, or $S^2 \times S^1$. These sections are metrically inhomogeneous because, intuitively, gravitational waves are rippling through space. But despite the inhomogeneities, these models have a two-parameter spacelike symmetry (technically, there are two commuting spacelike Killing vector fields). Chruściel et al. focus on polarized Gowdy spacetimes where the Killing fields are everywhere orthogonal. We know that initial data on a slice Σ determines a unique (up to isometry) maximal development for which Σ is a Cauchy surface. It is shown that for an open dense subset of this initial data, the maximal development is not properly extendible. In the $\Sigma = T^3$ case a generic solution is not extendible in the future because the solution evolves towards $t = +\infty$ with diminishing curvature. In the $\Sigma = S^3$ or $S^2 \times S^1$ cases a generic solution is not future extendible because it ends in a big crunch.

The supporters of cosmic censorship can also note that there is a respect in which the Oppenheimer–Synder–Volkoff model is suggestive of a feature which may be generic to gravitational collapse; namely, an event horizon forms whenever the collapse of matter (whose stress–energy–momentum tensor obeys appropriate energy conditions) proceeds beyond some critical stages such as the formation of a trapped surface. Even if true, this *event horizon conjecture* (EHC), as Werner Israel dubs it, does not by itself suffice to establish either the strong or the weak form of the CCH. Strong cosmic censorship may still fail because within the event horizon a locally naked singularity can develop. And weak cosmic censorship can fail because, as illustrated in Fig. 3.11, the formation of the event horizon may be preceded by the development of a globally naked singularity. Nevertheless, proving the EHC would be a big step towards proving the CCH. Israel (1986) has established a preliminary version of the EHC by assuming that a trapped surface develops and that its cylindrical extension remains non-singular.

A crucial test case for the CCH concerns vacuum solutions to EFE. For then worries about whether the coupled Einstein–matter equations admit a well-posed initial value problem vanish, as do worries about what conditions T^{ab} and the equations of state must satisfy in order to qualify as physically realistic. Some care is still needed in formulating the CCH in this setting. For example, the Taub–NUT spacetime falsifies the most naive attempt to formulate the CCH in this setting. *Conjecture 0*: Among vacuum solutions to Einstein field equations, future naked singularities (on any reasonable definition) do not develop from regular initial data.

One of two directions can be taken to maneuver around this counter-example. The first is to exclude by fiat acausal solutions, leading to *Conjecture 1*: Let M, g_{ab} be a maximally extended vacuum solution to Einstein's field equations; if $\Sigma \subset M$ is a partial Cauchy surface, then the spacetime is not future nakedly singular with respect to Σ unless strong causality is violated at some $p \in H^+(\Sigma)$. If 'future nakedly singular' is taken as FNS_1, then from the point of view of evolution we can formulate this conjecture as it applies to a compact initial data surface as follows: the maximal future Cauchy development of the appropriate initial data for a vacuum solution to Einstein's

there is an event horizon. Another reason for doubting the significance of the Shapiro–Teukolsky result comes from their use of the Vlasov equation to describe the distribution of dust particles in their model. Rendall (1992b) argued that it is unreasonable to expect that more general solutions for this equation will exhibit the singular behavior found in the Shapiro–Teukolsky model. (But in any case the use of computer models provides a powerful tool for the exploration of the cosmic censorship hypothesis, and given the difficulty in establishing interesting mathematical results it may be expected that much of the future progress in discovering counterexamples to various versions of cosmic censorship will be achieved with the help of this tool.)

Now let us turn the coin to look at the evidence in favor of cosmic censorship. Penrose (1973) noted that the CCH can be used to derive inequalities, involving areas of trapped surfaces and available masses, on the behavior of black holes.[27] He tried but failed to find physically plausible ways to violate these inequalities. The failure constitutes only weak evidence for cosmic censorship since, as Penrose himself notes, if cosmic censorship does not hold, a naked singularity may form without a trapped surface also forming. Wald's (1973) investigation starts with the inequality necessary for the Kerr–Newman solutions to represent a black hole: $M^2 \geq Q^2 + \mathcal{J}^2/M^2$, where M, Q, and $\mathcal{J}$ are respectively the mass, electric charge, and the angular momentum. He found that various ways of trying to violate this inequality by injecting charge and angular momentum into the black hole all fail. Again this gives only weak support to the CCH since although it confirms the stability of stationary black holes once they form, it gives no confidence that a black hole rather than a naked singularity will form in gravitational collapse.

Krolak (1986, 1987a, 1987b) has offered various censorship theorems, but they rely on the dubious assumption of the existence of marginally outgoing null geodesics.[28]

Newman (1984b) established a censorship theorem for conformally flat spacetimes. (M, g_{ab} is *conformally flat* just in case $g_{ab} = \Omega^2 \eta_{ab}$ where η_{ab} has vanishing Riemann curvature and Ω^2 is a C^∞ map from M to $(0, +\infty)$.) In a generic null convergent ($R_{ab}K^aK^b \geq 0$ for any null vector K^a) conformally flat spacetime all null geodesics are incomplete. Nevertheless, Newman showed that if Ω^2 is a proper map (i.e., the inverse image of any compact subset of $(0, +\infty)$ is a compact subset of M) then M, g_{ab} is not FNS_2. Of more potential significance is another result of Newman (1984a, 1986) that establishes a weak form of cosmic censorship for weakly asymptotically simple and empty spacetimes[29] which satisfy a suitable causality condition and in which every incomplete null geodesic experiences a persistent curvature of sufficient strength. However, the applicability of the persistent curvature condition to gravitational collapse and cosmology remains uncertain.

Chruściel, Isenberg, and Moncrief (1990) have established a form of strong cosmic censorship for a special class of Gowdy spacetimes. These are vacuum solutions to Einstein's field equations with vanishing cosmological constant. They are spatially closed with space sections Σ being topologically

T^3, S^3, or $S^2 \times S^1$. These sections are metrically inhomogeneous because, intuitively, gravitational waves are rippling through space. But despite the inhomogeneities, these models have a two-parameter spacelike symmetry (technically, there are two commuting spacelike Killing vector fields). Chruściel et al. focus on polarized Gowdy spacetimes where the Killing fields are everywhere orthogonal. We know that initial data on a slice Σ determines a unique (up to isometry) maximal development for which Σ is a Cauchy surface. It is shown that for an open dense subset of this initial data, the maximal development is not properly extendible. In the $\Sigma = T^3$ case a generic solution is not extendible in the future because the solution evolves towards $t = +\infty$ with diminishing curvature. In the $\Sigma = S^3$ or $S^2 \times S^1$ cases a generic solution is not future extendible because it ends in a big crunch.

The supporters of cosmic censorship can also note that there is a respect in which the Oppenheimer–Synder–Volkoff model is suggestive of a feature which may be generic to gravitational collapse; namely, an event horizon forms whenever the collapse of matter (whose stress–energy–momentum tensor obeys appropriate energy conditions) proceeds beyond some critical stages such as the formation of a trapped surface. Even if true, this *event horizon conjecture* (EHC), as Werner Israel dubs it, does not by itself suffice to establish either the strong or the weak form of the CCH. Strong cosmic censorship may still fail because within the event horizon a locally naked singularity can develop. And weak cosmic censorship can fail because, as illustrated in Fig. 3.11, the formation of the event horizon may be preceded by the development of a globally naked singularity. Nevertheless, proving the EHC would be a big step towards proving the CCH. Israel (1986) has established a preliminary version of the EHC by assuming that a trapped surface develops and that its cylindrical extension remains non-singular.

A crucial test case for the CCH concerns vacuum solutions to EFE. For then worries about whether the coupled Einstein–matter equations admit a well-posed initial value problem vanish, as do worries about what conditions T^{ab} and the equations of state must satisfy in order to qualify as physically realistic. Some care is still needed in formulating the CCH in this setting. For example, the Taub–NUT spacetime falsifies the most naive attempt to formulate the CCH in this setting. *Conjecture 0*: Among vacuum solutions to Einstein field equations, future naked singularities (on any reasonable definition) do not develop from regular initial data.

One of two directions can be taken to maneuver around this counter-example. The first is to exclude by fiat acausal solutions, leading to *Conjecture 1*: Let M, g_{ab} be a maximally extended vacuum solution to Einstein's field equations; if $\Sigma \subset M$ is a partial Cauchy surface, then the spacetime is not future nakedly singular with respect to Σ unless strong causality is violated at some $p \in H^+(\Sigma)$. If 'future nakedly singular' is taken as FNS_1, then from the point of view of evolution we can formulate this conjecture as it applies to a compact initial data surface as follows: the maximal future Cauchy development of the appropriate initial data for a vacuum solution to Einstein's

field equation prescribed for a compact Σ (without boundary) is not extendible as a solution of the field equations unless strong causality is violated at some $p \in H^+(\Sigma)$. (This follows from the same kind of argument used in section 3.3. As noted, in a globally hyperbolic spacetime any compact partial Cauchy surface is a Cauchy surface. So if cosmic censorship in the form of global hyperbolicity is to hold, $D^+(\Sigma)$ for a compact partial Cauchy Σ cannot be extendible.) The alternative to dismissing acausal solutions is to eschew spatially closed universes, leading to *Conjecture 2*: Let M, g_{ab} be a maximally extended vacuum solution to Einstein's field equations; if M, g_{ab} does not admit a compact partial Cauchy surface then it is not future nakedly singular with respect to any partial Cauchy surface.

This retreat from Conjecture 0 to Conjectures 1 and 2 is less than satisfying. Conjecture 1 rules out by fiat one sort of breakdown in predictability due to the emergence of acausal features from a past that may have been causally pure, while Conjecture 2 refuses to consider spatially closed universes. At present, however, more attractive alternatives do not seem to be available.

A failure of Conjecture 1 or 2 need not be seen as fatal to cosmic censorship since one could retreat to the version of censorship that asserts only that such failures are sparse among the vacuum field solutions. Genericity considerations can also be used to fill the gap in the above conjectures. Moving from Conjecture 0 to Conjecture 1 amounts to ignoring the violation of cosmic censorship that arises in Taub–NUT spacetime. That such ignorance may not be bliss was suggested by Moncrief's (1981) finding that there is an infinite dimensional family of vacuum solutions with NUT-type extensions. But Moncrief and Isenberg (1983) also established a result that suggests that such behavior is nevertheless special. For a partial Cauchy surface Σ in the Taub region of Taub–NUT spacetime, $H^+(\Sigma)$ is a compact null surface ruled by closed null curves. The result is that analytic vacuum solutions possessing such a null surface must also possess a Killing field.

It is not easy to say whether the weight of evidence to date favors some interesting form of cosmic censorship. Certainly there is enough evidence pro and con to keep both the proponents and opponents at work.

3.6 Black hole evaporation

If quantum gravity doesn't banish singularities altogether, then a frightening prospect opens up, even if cosmic censorship is true at the level of classical GTR. For according to calculations from semiclassical quantum gravity, black holes will eventually evaporate due to Hawking radiation (see Wald 1984a, Sec. 14.3). And assuming that classical relativistic spacetime concepts can be applied to the result, either a naked singularity or a thunderbolt can be expected to emerge. To discuss the reasons for that expectation, I will review a technical result due to Geroch and Wald (see Wald 1984b; see also Kodama 1979 for a related result).

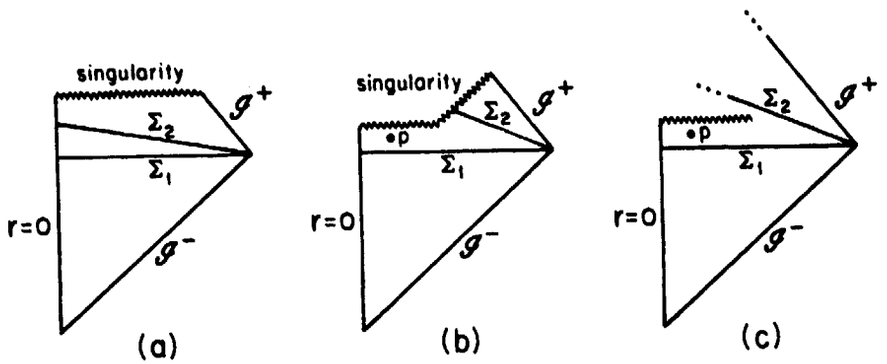

Fig. 3.12 Black hole evaporation

Theorem (Geroch and Wald). Let M, g_{ab} be a time-oriented spacetime, and let Σ_1 and Σ_2 be closed achronal sets with Σ_1 connected and Σ_2 edgeless. Suppose that (i) there is a point $p \in D^+(\Sigma_1)$ such that $p \notin (J^-(\Sigma_2) \cup J^+(\Sigma_2))$, and (ii) $J^+(K) \cap \Sigma_1$ has compact closure, where $K = \Sigma_1 - (D^-(\Sigma_2) \cap \Sigma_1)$. Then $\Sigma_2 \not\subset D^+(\Sigma_1)$.

The Penrose diagram of (one-half) of a generic black hole configuration of a gravitationally collapsed body with center of symmetry $r = 0$ is shown in Fig. 3.12a. In this case $D^+(\Sigma_1)$ for the time slice Σ_1 includes every point to the future of Σ_1 so that there is no slice Σ_2 to the future such that $\Sigma_2 \not\subset D^+(\Sigma_1)$. Here the conclusion of the theorem fails because neither of the conditions (i) or (ii) is applicable. Figure 3.12b pictures the evaporation of a black hole by means of a catastrophic burst of electromagnetic radiation (thunderbolt). No naked singularity develops since the burst of radiation cuts off future extensions in such a way that again $D^+(\Sigma_1)$ includes everything to the future of Σ_1.[30] The theorem fails to apply since although condition (i) holds, (ii) fails. However, if Hawking radiation does not produce thunderbolt evaporation, then something akin to Fig. 3.12c should result; since conditions (i) and (ii) both apply, the theorem can then be invoked to conclude that there will be a violation of cosmic censorship in the form of a breakdown in predictability. The question then becomes whether the calculations of semiclassical quantum gravity—which predict Hawking radiation—indicate that in black hole evaporation a naked singularity is more likely to result than a thunderbolt singularity. Hawking and Stewart (1993) and Lowe (1993) report somewhat conflicting results.

What I would like to briefly explore is the prospect of proving that black hole evaporation can produce a violation of cosmic censorship not only in the sense of a breakdown in predictability but also in the stronger sense of a singularity visible from $\mathcal{I}^+$, as Fig. 3.12c would suggest. Towards this end I will assume that black hole evaporation takes place in a spacetime where both $\mathcal{I}^+$ and $\mathcal{I}^-$ are defined. The theorem of Geroch and Wald already tells us

that when the hypotheses apply the future boundary $H^+(\Sigma_1)$ of $D^+(\Sigma_1)$ is non-empty. The generators of $H^+(\Sigma_1)$ are null geodesics. I will simply assume that these generators extend to $\mathscr{I}^+$. It may also be assumed without loss of generality that Σ_1 is a partial Cauchy surface and also that $\Sigma_1 \not\subset J^-(\mathscr{I}^+)$, for Σ_1 is supposed to correspond to a time before the black hole evaporates. Because Σ_1 is edgeless, the generators of $H^+(\Sigma_1)$ are past endless and, thus, past inextendible (Hawking and Ellis 1973, Prop. 6.5.3). There are now two main possibilities to consider. (1) Some of the generators of $H^+(\Sigma_1)$ are totally or partially past imprisoned in a compact set of the spacetime. This possibility can be ruled out by imposing the requirement of strong causality (Hawking and Ellis 1973, Prop. 6.4.7). (2) With (1) ruled out, the generators of $H^+(\Sigma_1)$ must in some sense "run off the edge" of spacetime in the past direction. There are two main subcases to consider. (a) Some of the generators run into a singularity, i.e., are past incomplete. Since, by assumption, the generators extend to $\mathscr{I}^+$, we have our naked singularity. To rest content with this subcase, we need to rule out the other. (b) The generators are all past complete. Again we have two subcases to consider. (α) The generators run into an "internal infinity," which can be illustrated by modifying the example accompanying Fig. 3.2 by letting Ω go to ∞ rather than to 0 as the region K is approached.[31] Imposing a suitable condition to rule out such pathologies we should arrive at the second subcase. (β) The generators of $H^+(\Sigma_1)$ all run to $\mathscr{I}^-$. But in this case we should be able to show that $\Sigma_1 \subset J^-(\mathscr{I}^+)$, which is contrary to assumption. In sum let me emphasize that there is no pretense of a proof here. But the considerations reviewed do seem to me to lend credence to the notion that, excluding thunderbolt evaporation, physically reasonable cases of black hole evaporation can be expected to produce singularities, in the sense of geodesic incompleteness, that are visible from $\mathscr{I}^+$, and thus are violations of WCC.

The upshot is potentially disturbing. If we believe the classical GTR, it is likely that black holes have formed throughout spacetime. If semiclassical quantum gravity has validity, then these black holes evaporate, leading eventually either to thunderbolt singularities or to naked singularities. If we are not always saved from naked singularities by thunderbolts, then perhaps we are saved by the fact that evaporation time for black holes as massive as our sun is very long indeed. But if mini-black holes formed in the early universe or if, as has been recently suggested, mini-black holes form in supernovas, we could be surrounded by naked singularities. Just how disturbing this prospect is cannot be assessed until we are in possession of more detailed information about the nature of the singularities that emerge from black hole evaporation.

3.7 What if cosmic censorship should fail?

How much of a disaster for physics would it be if the CCH should prove to be wrong? Early in the investigation of the problem of cosmic censorship,

Penrose posed this question and sketched a preliminary response:

> It is sometimes said that if naked singularities do occur, then this would be disastrous for physics. I do not share this view. We have already had the example of the big-bang singularity in the remote past, which seems not to be avoidable. The "disaster" to physics occurred right at the beginning. Surely the presence of naked singularities arising occasionally in collapse under more "controlled" circumstances would be the very reverse of a disaster. The effects of such singular occurrences could then be accessible to observation *now*. Theories of singularities would be open to observational test. The initial mystery of creation, therefore, would no longer be able to hide in the obscurity afforded by its supposed uniqueness. (Penrose 1973, p. 133)

The reference to the big bang singularity shows that the visibility of the singularity by itself portends no disaster; but this point has already been codified in the various definitions of naked singularities, which exclude the initial singularities in standard big bang cosmologies. Only those singularities that entail a breakdown in predictability and determinism are counted as naked. Now for whatever psychological reasons, we are less disturbed by an inability to retrodict or determine the past than by an inability to predict or determine the future. The effects of naked singularities in our past would imply a breakdown in retrodiction and historical determinism; but, as Penrose says, the effects of such singularities would be accessible to observation now, and given a knowledge of these effects we can hope to determine what will transpire in the future. This hope is undercut if our spacetime is future nakedly singular. Again we must ask how much of a disaster it would be for physics if we were exposed to such spacetime pathologies.

To make a start towards an answer it is necessary to sunder a potentially misleading association made above between predictability and determinism. Predictability in a generic relativistic spacetime is impossible since it is generally impossible to acquire enough information for a sure forecast before the occurrence of the event to be predicted (see chapter 5). However, determinism does hold at least locally for the pure gravitational field, and arguably it can be expected to hold locally for the coupled gravitational–matter fields for any matter fields that are fundamental (see section 3.4). The question of cosmic censorship is then whether this local determinism will break down in the large due to the development of naked singularities.

At this juncture the devil's advocate may propose that we should not get overly agitated by the prospect that a cosmic censor is not at work, for after all, QM has accustomed us to breakdowns in determinism. While giving the devil his due, it has to be noted that there are important differences between the indeterminism of QM and the indeterminism associated with a failure of cosmic censorship. In the quantum case the unitary evolution of the state vector, of which the Schrödinger equation is simply the infinitesimal form, is deterministic, and indeterminism enters only when the unitary evolution is

interrupted by a miracle of a "collapse of the state vector" when a measurement is made. The nature of this miracle and even whether it occurs are issues currently subject to intense debate. Whatever the outcome of the debate, the kind of indeterminism at issue is at worst not of the anything-goes variety since the quantum theory specifies the precise form for the statistics of outcomes of quantum measurements. By contrast, the principles of classical GTR do not tell us whether a naked singularity will passively absorb whatever falls into it or will regurgitate helter-skelter TV sets, green slime, or God only knows what. The fear here has been articulated by Shapiro and Teukolsky.

> A counterexample to cosmic censorship—a configuration of matter that evolves to produce a naked singularity—would be a catastrophe for the theory [of general relativity]. Given the existence of just one singularity, general relativity could not say anything precise about the future evolution of any region of space in communication with the singularity. Determining the validity of cosmic censorship is perhaps the most important outstanding problem in the study of general relativity.[32] (Shapiro and Teukolsky 1991b, p. 330)

Such fears about the unbridled influences of naked singularities would be somewhat assuaged if it could be shown that a naked singularity can have only a minimal influence on external observers. For example, one might hope that in the asymptotically flat regime such violations of WCC that do occur are mild in the sense that the null geodesics that emerge from the singularity and escape to $\mathscr{I}^+$ are of "measure zero" in some appropriate sense. In fact, however, this hope is not realized for the naked singularities that develop in self-similar and non-self-similar gravitational collapse of dust, perfect fluids, and pure radiation (see Joshi 1993, Ch. 6).

The further exploration of fears about the disruptive influence of naked singularities requires attention to technical questions that are not only hard to answer but also hard to formulate precisely. To give an indication of what sorts of questions need to be addressed, let us pursue the considerations raised in Approaches 4 and 5 of section 3.3. Choose a three-manifold Σ without boundary and specify initial data appropriate to the system of gravitational and matter fields one is interested in. Use the Einstein–matter equations to evolve the initial data in time as far as compatible with Σ being a Cauchy surface so that in the resulting spacetime M, g_{ab}, M is $D(\Sigma)$. (We know that there is a unique (up to diffeomorphism) way of doing this.) Now ask whether M, g_{ab} is extendible to a larger spacetime. We know that if Σ is non-compact the extendibility of M, g_{ab} may be due to an unfortunate choice of Σ and not to the appearance of a naked singularity. To avoid such worries we can concentrate on cases where we can be sure that the extendibility is not due to a bad choice of Σ (e.g., Σ is compact or else Σ is non-compact and Σ, h_{ab} is complete and asymptotically flat). A second hitch is that we are not interested in just any mathematical extension of $\tilde{M}, \tilde{g}_{ab}$ of M, g_{ab} but only in those that satisfy EFE. But this is not a well-defined notion until the T^{ab} is

specified for the extension, and it is far from clear what conditions, beyond the standard energy conditions, should be imposed on T^{ab} in the extended region. To avoid this difficulty for the moment, let us focus on the empty space initial value problem and on extensions that are also solutions to the vacuum EFE.

We are now in a position to formulate some more or less precise questions. Are there examples where there are inequivalent (i.e., non-isometric) extensions $\tilde{M}, \tilde{g}_{ab}$, $\tilde{\tilde{M}}', \tilde{\tilde{g}}_{ab}, \ldots$, of M, g_{ab}? If so, what is the multiplicity of such extensions? Can uniqueness of extensions be restored by requiring that the extensions preserve some features of M, g_{ab}, such as symmetries of g_{ab}? On the other hand, if there is a unique extension, is uniqueness undermined by restricting the allowable isometries, e.g., to ones that preserve orientation or to ones that do not move the original Cauchy surface Σ of the maximal globally hyperbolic development M, g_{ab}?[33] Of course, these questions are moot if there are no physically realistic examples where cosmic censorship fails. Not having any such examples in hand to investigate, we have to try to get a feel for what answers are likely to emerge by investigating mathematical examples where cosmic censorship fails.

It has been claimed that the maximal globally hyperbolic region of Taub–NUT spacetime (the Taub region) admits different extensions. The examples given in the literature do indeed exhibit the existence of different extensions, but these extensions are not inequivalent, i.e., non-isometric (see Chruściel and Isenberg 1993). But if the isometry is required to preserve a chosen Cauchy surface of the Taub region, then these extensions are inequivalent (Chruściel and Isenberg 1993). Furthermore, the existence of non-isometric extensions of the Taub region (without any restriction on the isometry) has been demonstrated by Chruściel and Isenberg (1993), although to achieve non-uniqueness the Taub region must be extended in the past as well as in the future. They also show that among the polarized Gowdy spacetimes that violate cosmic censorship there are some that admit of an infinite multiplicity of inequivalent extensions. These results suggest that if cosmic censorship fails for vacuum solutions then a variety of extensions across $H^+(\Sigma)$ will be available even if the extensions are restricted to vacuum solutions.

If naked singularities are not quiescent but throw off matter, then even if we start with a vacuum solution M, g_{ab} there is no a priori reason to limit the extension across $H^+(\Sigma)$ to vacuum solutions. So unless some limitations are placed on what a naked singularity can spew out, the specter of a vast additional range of underdetermination arises. Penrose has cautioned that not just anything goes here.

> If we envisage an isolated naked singularity as a source of new matter in the universe, then we do not *quite* have unlimited freedom in this! For although in the neighborhood of the singularity we have no equations, we still have normal physics in the space-time *surrounding* the singularity. From the mass–energy flux theorems of Bondi et al. and Sachs, it follows that it

> is *not* possible for more mass to be ejected from a singularity than the original total mass of the system, *unless* we are allowed to be left with a singularity of *negative* total mass. (Such a singularity would *repel* other bodies, but would still be attracted by them!) (Penrose 1969, p. 274)

Assuming that there is no naked singularity to begin with, the proof of the positive mass conjecture rules out the last awful possibility.[34] But although not just anything goes, the range of what a naked singularity can, consistently with standard GTR, disgorge is vast.

Another way to get a handle on possible non-uniqueness due to a failure of cosmic censorship is to suppose that M, g_{ab} has a unique extension across $H^+(\Sigma)$ or else that a reason has been found to single out one of the extensions. Treat the extended spacetime $\tilde{M}, \tilde{g}_{ab}$ as a fixed background spacetime, and ask whether test fields on the globally hyperbolic region of M, g_{ab} have unique extensions across $H^+(\Sigma)$. Only a few relevant results are known, and they are mixed. For elastic collisions of billiard balls in certain "wormhole spacetimes" where closed timelike curves develop when one crosses $H^+(\Sigma)$ it has been found that there is an infinite multiplicity of self-consistent but inequivalent extensions of some initial trajectories of a billiard ball that take the ball around a closed time loop and into collision with itself (see chapter 6).

On the other hand, Wald (1980) has argued that in static but non-globally hyperbolic spacetimes there is a physically reasonable way to obtain a global dynamics for a scalar field Φ obeying the homogeneous wave equation $\Box\Phi =: \nabla_a\nabla^a\Phi = 0$. If the value of Φ and its derivative normal to a spacelike Σ are specified, then the wave equation determines Φ only in $D(\Sigma)$, which in the case of the hypothesized non-globally hyperbolic spacetime is only a proper subset of the full spacetime. Wald's prescription fixes Φ throughout the spacetime in such a way to secure agreement with the usual dynamics on $D(\Sigma)$ and to make Φ smooth everywhere for smooth initial data of compact support. This is a nice result, but the prescription implicitly assumes boundary conditions on the singularities. For instance, in the artificial case where the singularities correspond to regions cut out of a larger singularity-free spacetime, Wald's prescription requires that Φ vanishes on the boundaries of the cutout regions. There are surely other prescriptions for obtaining global values which also agree with the normal dynamics on $D(\Sigma)$ and which make Φ smooth everywhere for smooth initial data. Adjudicating among the alternative prescriptions will depend on which of the implicit boundary conditions on singularities are regarded as reasonable. It is not clear what principles can be used to guide these decisions.

Friedman and Morris (1991a, b) have demonstrated the uniqueness of solutions of $\Box\Phi = 0$ for a class of non-globally hyperbolic spacetimes that contain closed timelike curves. However, these spacetimes do not conform to the paradigm studied in this chapter since they contain no partial Cauchy surfaces and initial data have to be specified on $\mathscr{I}^-$.

Now for sake of discussion let us be pessimistic. Suppose that cosmic censorship fails and that as a consequence there is a vast range of alternative extensions across the Cauchy horizon, all of which are compatible with principles of classical GTR. Must we conclude that physics becomes hopeless? Must we simply throw up our hands and wait for the weirdness to unfold? No! We can try to discern what regularities naked singularities display. For example, are the singularities that develop in certain situations quiescent? Do those that develop in other situations all ooze green slime, and if so do they ooze it at a regular rate? The attitude that physics is hopeless if naked singularities occur stems from what may be termed GTR chauvinism—the notion that Einstein and his followers discovered all of the laws relevant to classical gravitation. If we acknowledge that laws of nature are simply codifications of certain deep regularities,[35] then we should be prepared to discover through observation that naked singularities obey laws of their own. If we are lucky these additional laws, when conjoined with the laws of standard GTR, will restore predictability and determinism. Even if we are not so lucky they may still give us some interesting physics. Of course, we must be prepared for the eventuality that naked singularities exhibit no interesting regularities at all, in which case they would indeed be a disaster for physics. But at present only GTR chauvinism would lead us to fixate on this worst-case scenario.

3.8 A dirty open secret

Cosmic censorship as it has been investigated here is concerned primarily with determinism. But it is "well known" (to the handful of experts who think about these matters) that determinism in general relativity fails without the help of what looks very much like fiat.

For concreteness and simplicity, consider solutions to the source-free EFE. Pick any two such solutions M, g_{ab} and M', g'_{ab} that are maximal and choose any spacelike hypersurfaces $\Sigma \subset M$ and $\Sigma' \subset M'$. The metrics of the respective spacetimes induce on the hypersurfaces their first and second fundamental forms, resulting in the initial data sets Σ, h_{ab}, K_{ab} and $\Sigma', h'_{ab}, K'_{ab}$. One would like to be able to show that if these initial data sets agree (i.e., there is a diffeomorphism d of Σ onto Σ' such that $d^*h_{ab} = h'_{ab}$ $d^*K_{ab} = K'_{ab}$) then so do the respective domains of dependence $D(\Sigma) \subseteq M$ and $D(\Sigma') \subseteq M'$ (i.e., there is an extension of d to an isometry of $D(\Sigma)$ onto $D(\Sigma')$). Having shown this, one can then go on to the question of whether for appropriate Σ and Σ', $D(\Sigma)$ and $D(\Sigma')$ exhaust M and M' respectively and, if not, whether $M - D(\Sigma)$ and $M' - D(\Sigma')$ agree—i.e., the question of cosmic censorship as studied under Approaches 4 and 5 of section 3.3. But the previous problem must be attended to first, for if it should have a negative outcome, the failure of determinism involved precedes any lapse by the cosmic censor.

The problem might seem to have an obviously positive outcome. The theorem of Choquet-Bruhat and Geroch (1969) proves the existence of a

unique (up to isometry) maximal development for which the initial value hypersurface is a Cauchy surface. So if each of the spacetimes considered is such that for any spacelike hypersurface, the associated maximal development is realized in the spacetime, then we are done. And it might seem that the antecedent is discharged by the already adopted demand that the spacetimes under consideration are maximal.

Unfortunately, the last step does not hold. That a spacetime M, g_{ab} is maximal does not guarantee that it is *determinism maximal* in the sense that there is no spacelike $\Sigma \subset M$ such that an isometric imbedding $\varphi: D(\Sigma) \to \tilde{M}$ into a spacetime $\tilde{M}, \tilde{g}_{ab}$ can be found that makes $\varphi(D(\Sigma))$ a proper subset of $D(\varphi(\Sigma))$.[36] To create an example of a maximal but not determinism maximal spacetime, start with a spacetime just as nice as you like—say, four-dimensional Minkowski spacetime. Remove the two-plane $\{t = x = 1\}$. In the resulting holey spacetime a time slice such as $\Sigma = \{t = 0\}$ is no longer a Cauchy surface. Of course, the holey spacetime is properly extendible and is thereby excluded from consideration by the initial stipulation. However, the universal covering spacetime of the holey spacetime is maximal. But it is not determinism maximal since if $\tilde{\Sigma}$ is any of the lifts of Σ to the universal cover, there is an isometric imbedding φ of $D(\tilde{\Sigma})$ back into Minkowski spacetime such that $\varphi(D(\tilde{\Sigma})) \subset D(\varphi(\tilde{\Sigma}))$ (see Geroch 1977). Such examples cannot be brushed aside as being rare since for any normal spacetime one can create innumerable mutilated versions that contain determinism holes.

Determinism can be restored by declaring any spacetime within which maximal developments do not get realized as persona non grata. But on what grounds that are not question begging? Certainly not on the grounds of failure to satisfy the field equations and energy conditions. To rule out the above example is to rule out one way Nature might, consistently with all of the known laws of GTR, continue to evolve things across $H^+(\Sigma)$. What then is to say that She cannot proceed this way? The most prevalent attitude among general relativists seems to be that fiat is required (see Ellis and Schmidt 1977), otherwise questions about more interesting ways in which determinism can fail are never reached. I implicitly adopted this attitude in the foregoing sections. I am not proud of doing so, but I am no better than my brethren in physics in seeing an alternative to fiat.

3.9 Conclusion

In chapter 2 I argued that questions about the existence of singularities may turn on delicate questions about the continuity and differentiability of the metric. Questions of cosmic censorship also involve delicacies of continuity and differentiability. In the present chapter I ignored them for the most part, for to have done otherwise would have turned an already complicated discussion into an unreadable one. But ultimately these considerations must be incorporated. Consider, for example, a theorem which purports to

demonstrate some form of cosmic censorship by showing that the maximal Cauchy development of some generic set of initial data is inextendible. One must ask: What are the differentiability conditions assumed in the theorem? Are they low enough to assure that all physically reasonable extensions have been taken into account? One plausible attitude here is that, for purposes of evaluating cosmic censorship, one may assume that the spacetime metric has enough differentiability to assure the existence and uniqueness of Cauchy developments, for otherwise determinism could fail even before the matter gets submitted to a censor. Sufficient conditions to assure this existence and uniqueness are given in Hawking and Ellis (1973, pp. 249–251), but minimally sufficient conditions are apparently not known. An opposing attitude would allow that once maximal Cauchy evolution stops, Nature might continue in some lower differentiable way as long as EFE are satisfied in some distributional sense.[37] If the latter attitude is accepted, the censor has to work much harder.

Even leaving aside such technicalities, the prospects seem dim for a quick and clean resolution of the question of cosmic censorship. The above discussion reveals that "cosmic censorship" is a label for a large and diverse class of ideas and motivations. It is doubtful that these various ideas and motivations can be accomodated in a few precise mathematical conjectures that can then be either proved or refuted. For both the proponents and opponents of cosmic censorship this means that there is no easy road to victory. The proponents have to be content with proving limited censorship theorems for special cases; the opponents have to pile up more and more counterexamples to various forms of cosmic censorship; and each side has to hope that eventually it will wear the other down. It is too soon to try to predict the outcome of this battle of attrition, but two things seem certain. First, whatever the final resolution, a good deal of interesting physics and interesting philosophy of physics will be generated by the process. Second, the battle is not apt to fizzle out for lack of interest, for the business of cosmic censorship is too important to ignore.

Notes

1. Although I will use predictability and determinism interchangeably in the next few sections, it will eventually become important to distinguish them; see section 3.7 and Earman (1986).

2. It may be that initial conditions on Σ plus supplementary boundary conditions may suffice for a determination; see section 3.7.

3. More precisely, $H^+(\Sigma)$ is defined as $D^+(\Sigma) - I^-(D^+(\Sigma))$. The *past Cauchy horizon* $H^-(\Sigma)$ of Σ is defined analogously.

4. It is assumed throughout that the spacetimes under discussion are *temporally orientable* so that they admit a globally consistent time directionality. For a definition of temporal orientability and a discussion of the hierarchy of causality conditions on relativistic spacetimes, see chapter 6.

5. The Penrose diagrams bring infinity in to a finite distance by means of a conformal transformation that preserves causal relations; see Hawking and Ellis (1973) for details. In Fig. 3.3b i^0 denotes spatial infinity.

6. Future null infinity $\mathscr{I}^+$ and past null infinity $\mathscr{I}^-$ are defined in a non-physical spacetime obtained by taking the conformal completion of physical spacetime (the reader may consult Hawking and Ellis (1973) and Wald (1984a) for details). The interior of a black hole B and the absolute event horizon E (see below) are defined in terms of $\mathscr{I}^-$, which lives in the non-physical spacetime. But B and E are part of physical spacetime.

7. In Fig. 3.4b i^+ and i^- denote respectively *future* and *past timelike infinity*. They are respectively the terminus and origin of timelike geodesics.

8. Roughly, a *stably causal* spacetime is one in which it is possible to widen the light cones without permitting closed causal loops to form; see chapter 6 for a precise definition.

9. Roger Penrose (1979) has put forward the "Weyl curvature hypothesis" which rules out the type of white holes that violate cosmic censorship. He conjectures that this asymmetry must be underwritten by a fundamental law of physics that is not time reversal invariant; see chapter 5.

10. See Gowdy (1977). The reader is invited to construct an example to show why the definition of topology change must be restricted to maximal partial Cauchy surfaces.

11. $H^+(\Sigma)$ for a spacelike hypersurface Σ is a null surface and its generators are null geodesics. The generators are either past endless or else have a past endpoint in the edge of Σ. For a partial Cauchy surface, which by definition is edgeless, the generators are past endless.

12. See Penrose (1978) for definitions of these concepts and for a discussion of how these concepts can be used to formulate cosmic censorship.

13. If $\mathfrak{R} \neq \varnothing$ because there is a $p \in M$ such that $I^-(p)$ contains a future-directed timelike half-curve γ of infinite proper length, then we have an example of *Malament–Hogarth spacetime*. These spacetimes are investigated in chapter 4.

14. In Droste coordinates the line element has the form

$$ds^2 = \left(1 - \frac{\alpha}{r}\right)^{-1} dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) - \left(1 - \frac{\alpha}{r}\right) dt^2$$

where α is proportional to the total mass M. If $\alpha < 0$ the metric components are regular down to $r = 0$ where a naked singularity resides.

15. Recall that, roughly speaking, strong causality rules out closed and almost closed causal curves.

16. The *constraint equations* are

$$^3R - K^{ab}K_{ab} + (K^a{}_a)^2 = 0$$

$$D_b K^{ab} - D_a K^b{}_b = 0$$

where ^3R is the curvature scalar of the metric h_{ab} of Σ and D_a is the derivative operator associated with h_{ab}. The second fundamental form (aka *extrinsic curvature*) K_{ab} of Σ is defined as follows. Let ξ^a be the normed tangent field of the congruence of timelike geodesics orthogonal to Σ. Then $K_{ab} =: \nabla_a \xi_b$. It can be shown that $K_{ab} = (1/2)\pounds_\xi h_{ab}$ where $\pounds_\xi$ is the Lie derivative with respect to ξ^a (see Wald 1984a, Appendix C). In a coordinate system x^α, t, where the x^α, $\alpha = 1, 2, 3$, label points on Σ and t is a parameter along the geodesics orthogonal to Σ such that $\xi^a = (\partial/\partial t)^a$, $K_{ab} = (1/2)\, \partial h_{ab}/\partial t$.

17. M, g_{ab} is a *development* of Σ, h_{ab}, K_{ab} if there is an imbedding $\varphi: \Sigma \to M$ such that $\varphi(\Sigma)$ is a spacelike hypersurface whose first and second fundamental forms induced by g_{ab} are respectively h_{ab} and K_{ab}.

18. Here I am using the terminology of Israel (1984).

19. Here I am following Geroch and Horowitz (1979).

20. Choosing Ω to go to ∞ as K is approached would also produce a naked singularity in Penrose's sense. But this naked singularity will not be a singularity in the originally intended sense, e.g., involving geodesic incompleteness (see chapter 2).

21. The terms *strong energy condition* and *weak energy condition* are potentially misleading since the former does not entail the latter. The dominant energy condition does entail the weak energy condition. The weak energy condition entails the *null energy condition* which requires that $T_{ab}K^aK^b \geq 0$ for every null vector K^a.

22. See however the discussion of acausal spacetimes in chapter 6.

23. A spacetime is *self-similar* just in case there is a timelike vector field V_a such that $\nabla_a V_b - \nabla_b V_a = \Theta g_{ab}$, Θ = constant. For a spherically symmetric spacetime, this means that under a coordinate transformation $t \to \bar{t} = at$, $r \to \bar{r} = ar$, $\theta \to \bar{\theta} = \theta$, $\phi \to \bar{\phi} = \phi$, the metric components g_{ij} $(i, j = 1, 2, 3, 4)$ transform as $\bar{g}_{ij}(\bar{r}, \bar{t}) = (1/a^2)g_{ij}(\bar{r}, \bar{t})$.

24. The description of a source as a perfect fluid involves idealizations. The fact that it is difficult to obtain existence and uniqueness theorems for the coupled Einstein–Euler equations when the fluid has compact spatial support (see Rendall 1992a) would, according to the present faith, not undermine determinism but would only serve to indicate the limits of the usefulness of the perfect fluid description.

25. The findings of Shapiro and Teukolsky are in accord with Kip Thorne's *hoop conjecture* which posits that in the non-spherical collapse of a massive object, a black hole forms if and only if the circumference of the object in all directions is less than its critical circumference $4\pi GM/c^2$. If correct, the hoop conjecture supports a limited form of cosmic censorship but would also point to a class of counterexamples to the general cosmic censorship hypothesis.

26. An *apparent horizon* is the outer boundary of a trapped region (see Hawking and Ellis 1973, Sec. 9.2 for precise definitions). The relevant point here is that the existence of an apparent horizon entails the existence of an event horizon but not conversely.

27. Recall that a *trapped surface* $\mathfrak{T}$ is a closed two-surface such that both ingoing and outgoing null geodesics orthogonal to $\mathfrak{T}$ are converging.

28. A null geodesic of M, g_{ab} that forms an achronal set and has an endpoint in $\mathscr{I}^+$ is called *outgoing*. It is *marginally outgoing* if it is a limit curve of all outgoing null geodesics.

29. See Hawking and Ellis (1973, p. 252) for a definition of this class of spacetimes.

30. The thunderbolt singularity could also spread to $\mathscr{I}^+$ along a spacelike path and cosmic censorship would be preserved.

31. The notion that M, g_{ab} contains an *internal infinity* can perhaps be captured by the condition that there is a neighborhood $U \subset M$ which has a compact boundary and which contains a geodesic half-curve of infinite affine length. I am indebted to Al Janis for this suggestion.

32. Given Einstein's aversion to the notion that God plays dice with the world, it is tempting to speculate that a nascent version of such concerns was behind his visceral dislike of spacetime singularities. However, Einstein showed no inclination to

distinguish between "good singularities," with which we can peacefully coexist, and "bad singularities," with which even detente is impossible. He thought that all singularities must be excluded from a complete scientific theory (see chapter 1).

33. Let $\psi: \Sigma \to M$ denote the imbedding of Σ as a Cauchy surface of M, g_{ab} and let $\Psi: M \to \tilde{M}$ and $\tilde{\Psi}: M \to \tilde{\tilde{M}}$ denote the isometric imbeddings of M, g_{ab} into $\tilde{M}$, $\tilde{g}_{ab}$ and $\tilde{\tilde{M}}$, $\tilde{\tilde{g}}_{ab}$ respectively. Chruściel and Isenberg (1993) note that if the Cauchy surface $\psi(\Sigma)$ has a privileged status, then one might not want to count $\tilde{M}$, $\tilde{g}_{ab}$ and $\tilde{\tilde{M}}$, $\tilde{\tilde{g}}_{ab}$ as equivalent under an isometry $\varphi: \tilde{M} \to \tilde{\tilde{M}}$ if φ *moves* Σ in the sense that $\varphi \circ \Psi \circ \psi \neq \tilde{\Psi} \circ \psi$.

34. See Wald (1984a) for a discussion of the positive mass proof. Negative mass Schwarzschild spacetime (see note 14) has negative ADM mass and a naked singularity. It is not a counterexample to the positive mass theorem, which requires that the initial hypersurface be singularity free. Nor, as noted above, would this example be regarded as a violation of the form of cosmic censorship that excludes only those naked singularities that develop from regular initial data, since the singularity has been present for all times.

35. See chapter 6 for a discussion of the nature of physical laws.

36. A determinism maximal spacetime is also known in the literature as a *hole-free* spacetime. Being hole free does not entail being globally hyperbolic (think of Gödel spacetime). Nor does the implication go in the opposite direction (think of a truncated version of Minkowski spacetime with all of the points such that $t \geq 1997$ removed). However, a globally hyperbolic and inextendible spacetime is necessarily hole free.

37. See chapter 2 for a discussion of this matter.

4

Supertasks

4.1 Introduction

Is it possible to perform a supertask, that is, to carry out an infinite number of operations in a finite span of time? In one sense the answer is obviously yes since, for example, an ordinary walk from point a to point b involves crossing an infinite number of finite (but rapidly shrinking) spatial intervals in a finite time. Providing a criterion to separate such uninteresting supertasks from the more interesting but controversial forms is in itself no easy task,[1] but there is no difficulty in providing exemplars of what philosophers have in mind by the latter. There is, for instance, the Thomson lamp (Thomson 1954–55). At $t = 0$ the lamp is on. Between $t = 0$ and $t = 1/2$ the switch at the base of the lamp is pressed, turning the lamp off. Between $t = 1/2$ and $t = 3/4$ the switch is pressed again, turning the lamp on. Etcetera. The upshot is that an infinite number of presses are completed by $t = 1$. Then there is the super π machine. Between $t = 0$ and $t = 1/2$ it prints the first digit of the decimal expansion of π. Between $t = 1/2$ and $t = 3/4$ it prints the second digit. Etcetera. The result is that the complete expansion has been printed at $t = 1$. More interestingly from the point of view of mathematical knowledge there is the Plato machine which checks some unresolved existential conjecture of number theory for '1' during the first half-second, for '2' during the next quarter-second. Etcetera. The result is that the truth value of the conjecture is determined at the end of one second.

Thomson thought that such devices are logically or conceptually impossible. The operation of the Thomson lamp (a misnomer if Thomson were correct) entails that (i) for any t such that $0 < t < 1$, if the lamp is off at t, then there is a t' such that $t < t' < 1$ and the lamp is on at t', and (ii) for any t such that $0 < t < 1$, if the lamp is on at t, then there is a t' such that $t < t' < 1$ and the lamp is off at t'. Thomson thought that it followed from (i) that the lamp cannot be off at $t = 1$ and from (ii) that the lamp cannot be on at $t = 1$, a contradiction since it is assumed that the lamp must be in one or the other of these states at any instant. The fallaciousness of the argument was pointed out by Benacerraf (1962).

Others have held that though conceptually possible such devices are

physically impossible. Benacerraf and Putnam (1983, p. 20), for example, seem to have thought that these devices are kinematically impossible due to the fact that relativity theory sets c (the velocity of light) as the limit with which the parts of the device can move. Again, however, the impossibility is not as obvious as claimed. A demonstration is needed to rule out as a kinematic possibility that the operation of the device is arranged so that with each successive step the distance the parts have to move (as in an ordinary stroll from a to b) shrinks sufficiently fast that the bound c is never violated. Of course, even if the device can be shown to pass muster at the kinematic level, it may still fail to satisfy necessary conditions for a dynamically possible process (see Grünbaum 1968, 1969 for a discussion).

I have nothing new to add to this discussion here.[2] Rather, my focus will be on the ways that the relativistic nature of spacetime can be exploited so as to finesse the accomplishment of a supertask. Very crudely, the strategy is to use a division of labor. One observer has available to her an infinite amount of proper time, thus allowing her to carry out an infinite task in an unremarkable way. For example, she may check an unresolved conjecture in number theory by checking it for '1' on day one, for '2' on day two, etc., ad infinitum. (Or if, as the numerals increase, she needs increasing amounts of time to complete the check, she can allow herself $f(n)$ days to check the conjecture for 'n', where $f(n)$ is any increasing function of n as long as $f(n) < \infty$ for all n.) A second observer, who uses up only a finite amount of his proper time, is so situated that his past light cone contains the entire world line of the first observer. The second observer thus has direct causal access to the infinite computation of the first observer, and in this way he obtains knowledge of the truth value of the conjecture in a finite amount of time. If this is genuinely possible in relativity theory, there is an irony involved. Prima facie relativity might have been thought to make supertasks more difficult if not impossible by imposing kinematic limitations on the workings of Thomson lamps, Plato machines, and the like. But on further analysis relativity theory seems to open up a royal road that leads to the functional equivalent of the accomplishment of a supertask. The rough sketch just given contains an unjustified optimism. We will see that relativistic spacetimes do provide opportunities for carrying out the functional equivalents of supertasks, but we will also see that they do so at a price. One approach is to set the supertask in a well-behaved spacetime (see section 4.2). Here a double price has to be paid; for the second observer who tries to take advantage of the infinite labor of the first observer must submit himself to unbounded forces that end his existence, and in any case he never observes the completion of the infinite labor at any definite time in his existence.

Alternatively, both of these difficulties can be overcome by exploiting spacetimes with unusual structures which I will dub *Malament–Hogarth spacetimes*. A large part of this chapter will be devoted to articulating the senses in which these spacetimes are physically problematic. As Hogarth has already shown, they are not globally hyperbolic (Lemma 4.1, section 4.3), so

that they violate strong cosmic censorship. And they may also violate other requirements one would expect a physically realistic spacetime to fulfill (section 4.6). It will turn out that the failure of global hyperbolicity occurs in a way which necessarily defeats attempts to control disturbances to the signaling between the first and second observer from singularities and other sources. This signaling will prove to be problematic in other ways. It may demand that the second observer pursue his own mini-supertask in his neighborhood of spacetime, forfeiting the advantage that a Malament–Hogarth spacetime was supposed to offer (section 4.7). Again, the signaling will be associated with indefinite blueshifts (Lemma 4.2, section 4.5), so that the energy of the signals can be indefinitely amplified, threatening to destroy the second observer who receives them.

4.2 Pitowsky spacetimes

The first published attempt to make precise the vague ideas sketched in section 4.1 for using relativistic effects to finesse supertasks was that of Pitowsky (1990). His approach is encapsulated in the following definition.

DEFINITION 4.1
M, g_{ab} is a *Pitowsky spacetime* just in case there are future-directed timelike half-curves $\gamma_1, \gamma_2 \in M$ such that $\int_{\gamma_1} d\tau = \infty$, $\int_{\gamma_2} d\tau < \infty$, and $\gamma_1 \subset I^-(\gamma_2)$.

The blandest relativistic spacetime of all, Minkowski spacetime, is Pitowskian, as shown by Pitowsky's own example. (It seems a safe conjecture that this example can be generalized to show that any relativistic spacetime that possesses a timelike half-curve of infinite proper length is Pitowskian.) Choose an inertial coordinate system $(\mathbf{x}, t)$. Let γ_1 be the timelike half-geodesic $\mathbf{x}(t) = \text{constant}$, $0 \leq t < +\infty$. Choose γ_2 to be a timelike half-curve that spirals around γ_1 in such a way that it keeps γ_1 in its causal shadow and that its tangential speed is $u(t) = [1 - \exp(-2t)]^{1/2}$, $c = 1$. The proper time for γ_2 is $d\tau = \exp(-t)\, dt$, so that $\int_{\gamma_2} d\tau = 1$. Those familiar with the "twin paradox" may wish to take this example as the extreme case of the paradox with γ_2 as the ultimate traveling twin who ages biologically only a finite amount while his stay-behind twin ages an infinite amount. But admittedly this example does not conform to the usual twin paradox scenario where the twins hold a final reunion.

Pitowsky tells the following story about this example.

> While [the mathematician] M [γ_2] peacefully cruises in orbit, his graduate students examine Fermat's conjecture one case after the other. . . . When they grow old, or become professors, they transmit the holy task to their own disciples, and so on. If a counterexample to Fermat's conjecture is ever encountered, a message is sent to [M]. In this case M has a fraction of a second to hit the brakes and return home. If no message arrives, M

> disintegrates with a smile, knowing that Fermat was right after all. (Pitowsky 1990, p. 83)

(The example is now somewhat dated since a proof of Fermat's last theorem has been offered. However, some lingering doubts may remain since the purported proof is over 200 pages. In any case, the punch of Pitowsky's story can be preserved by substituting for Fermat's theorem any unresolved conjecture of number theory with a prenex normal form consisting either of all universal quantifiers or else all existential quantifiers. Or the logician may wish to contemplate the problem of deciding for a formal system strong enough for arithmetic whether or not a given well-formed formula is a theorem.)

There are two things wrong with this story. The first concerns the notion that "M [γ_2] cruises peacefully in orbit." For ease of computation, assume that the mathematician γ_2 undergoes linear acceleration with $u(t)$ as before. The magnitude of acceleration $a(t) =: (A_b(t)A^b(t))^{1/2}$, where A^b is the four-vector acceleration, is $\exp(t)/[1 - \exp(-2t)]^{1/2}$, which blows up rapidly. (To stay within a linearly accelerating γ_2's causal shadow, γ_1 would also need to accelerate. But γ_1's acceleration can remain bounded. Indeed, γ_1 can undergo constant ("Born") acceleration, which guarantees that γ_1's velocity approaches the speed of light sufficiently slowly that its proper length is infinite.) Thus, any physically realistic embodiment of the mathematician will be quickly crushed by g-forces. The mathematician disintegrates with a grimace, perhaps before learning the truth about Fermat's conjecture. What is true in this example is true in general since any ultimate traveling twin in Minkowski spacetime must have unbounded acceleration. If the ultimate traveling twin moves rectilinearly and has an upper bound to his acceleration, then another traveler, Born accelerated at this upper bound, would achieve equal or greater velocity at each instant and therefore age less. But this Born accelerated traveler's world line has infinite proper length. Therefore the rectilinearly accelerated traveller must have no upper bound to his acceleration if he is to have finite total proper time. This result holds a fortiori for the general case of a traveler in curvilinear motion, for part of his acceleration will be transverse to the direction of motion, thus generating no velocity over time and no resultant clock slowing.

The second and conceptually more important difficulty with Pitowsky's story concerns the claim that the mathematician γ_2 can use the described procedure to gain sure knowledge of the truth value of Fermat's conjecture. If Fermat was wrong, γ_2 will eventually receive a signal from γ_1 announcing that a counterexample has been found, and at that moment γ_2 knows that Fermat was wrong. On the other hand, if Fermat was right, γ_2 never receives a signal from γ_1. But at no instant does γ_2 know whether the absence of a signal is because Fermat was right or because γ_1 has not yet arrived at a counterexample. Thus, at no definite moment in his existence does γ_2 know that Fermat was right. The fictitious mathematical sum of all of γ_2's stages

knows the truth of the matter. But this is cold comfort to the actual, non-mathematical γ_2. By way of analogy, if your world line γ is a timelike geodesic in Minkowski spacetime and you have drunk so deep from the fountain of youth that you live forever, then $I^-(\gamma)$ is the entirety of Minkowski spacetime. So the fictitious sum of every stage of you can have direct causal knowledge of every event in spacetime. But at no definite time does the actual you possess such global knowledge.

4.3 Malament–Hogarth spacetimes

Malament (1988) and Hogarth (1992) sought to solve the conceptual problem with Pitowsky's example by utilizing a different spacetime structure.

DEFINITION 4.2
M, g_{ab} is a *Malament–Hogarth spacetime* just in case there is a timelike half-curve $\gamma_1 \subset M$ and a point $p \in M$ such that $\int_{\gamma_1} d\tau = \infty$ and $\gamma_1 \subset I^-(p)$.

This definition contains no reference to a receiver γ_2. But if M, g_{ab} is a Malament–Hogarth (hereafter, M–H) spacetime, then there will be a future-directed timelike curve γ_2 from a point $q \in I^-(p)$ to p such that $\int_{\gamma_2(q,p)} d\tau < \infty$, where q can be chosen to lie in the causal future of the past endpoint of γ_1. Thus, if γ_1 proceeds as before to check Fermat's conjecture, γ_2 can know for sure at event p that if he has received no signal from γ_1 announcing a counterexample, then Fermat was right.

Such arrangements can also be used to "effectively decide" membership in a recursively enumerable but non-recursive set of integers.[3] To decide whether or not a given n is a member, γ_1 proceeds to enumerate the members of the set. By assumption, this can be done effectively. As each new member is generated, γ_1 checks to see whether it is equal to n. This too can be done effectively. γ_1 sends a signal to γ_2 just in case she gets a match. Consequently, γ_2 knows that n is a member precisely if he has received a signal by the time of the M–H event.

These scenarios cannot be carried out in Minkowski spacetime, as follows from

LEMMA 4.1 *An M–H spacetime is not globally hyperbolic.*

A formal proof of Lemma 4.1 was given by Hogarth (1992). A simple informal proof follows from the facts that a globally hyperbolic spacetime M, g_{ab} contains a Cauchy surface and that a spacetime with a Cauchy surface can be partitioned by a family of Cauchy surfaces. Suppose for purposes of contradiction that M, g_{ab} is both globally hyperbolic and contains an M–H point $p \in M$, i.e., that there is a future-directed timelike half-curve γ such that $\gamma \subset I^-(p)$ and $\int_\gamma d\tau = \infty$. Choose a Cauchy surface Σ through p, and extend

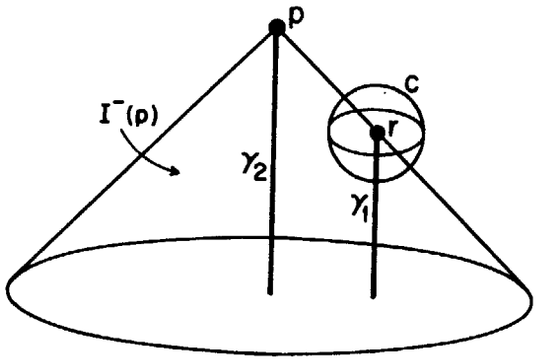

Fig. 4.1 A toy Malament–Hogarth spacetime

γ maximally in the past. This extended γ' is also contained in $I^-(p)$. Since γ' has no past or future endpoint, it must intersect Σ. But then since there is a timelike curve from the intersection point to p, Σ is not achronal and cannot, contrary to assumption, be a Cauchy surface.[4]

What of the problem in Pitowsky's original example that the receiver γ_2 has to undergo unbounded acceleration? In principle, both γ_1 and γ_2 can be timelike geodesics in at least some M–H spacetimes. The following toy example illustrates the point and also serves as a useful concrete example of an M–H spacetime. Start with Minkowski spacetime $\mathbf{R}^4, \eta_{ab}$ and choose a scalar field Ω which has value 1 outside of a compact set C (see Fig. 4.1) and which goes rapidly to $+\infty$ as the point r is approached. The M–H spacetime is then M, g_{ab} where $M = \mathbf{R}^4 - r$ and $g_{ab} = \Omega^2\eta_{ab}$. Timelike geodesics of η_{ab} in general do not remain geodesics in g_{ab}, but Ω can be chosen so that γ_1 is a geodesic of g_{ab} (e.g., if γ_1 is a geodesic of η_{ab}, choose an Ω with γ_1 as an axis of symmetry).

4.4 Paradoxes regained?

Consider again the super π machine which is supposed to print all the digits in the decimal expansion of π within a finite time span. Even leaving aside worries about whether the movement of the parts of the machine can be made to satisfy obvious kinematic and dynamic requirements, Chihara (1965) averred that there is something unintelligible about this hypothetical machine.

> The difficulty, as I see it, is not insufficiency of time, tape, ink, speed, strength or material power, and the like, but rather the inconceivability of how the machine could actually finish its supertask. The machine would supposedly print the digits on tape, one after another, while the tape flows through the machine, say from right to left. Hence, at each stage in the calculation, the sequence of digits will extend to the left with the last digit printed being "at

> center." Now when the machine completes its task and shuts itself off, we should be able to look at the tape to see what digit was printed last. But if the machine finishes printing all the digits which constitute the decimal expansion pi, no digit can be the last digit printed. And how are we to understand this situation? (Chihara 1965, p. 80)

Note first that the baldest form of Chihara's worry does not apply to the setup that has been imagined for M–H spacetimes; for the tape will not be available for γ_2's inspection since γ_1 goes crashing into a singularity or disappears to infinity. However, it might seem that a more sophisticated version of Chihara's conundrum can be mapped onto the M–H set up as follows. γ_1, who has available to her an infinite amount of proper time, prints the digits of π, say, one per second. And at the end of each step she sends a light signal to γ_2 announcing the result. γ_2 has a receiver equipped with an indicator which displays 'even' or 'odd' according as the case may be. By construction there is a $p \in \gamma_2$ at which γ_2 has received all of the signals from γ_1. One can then ask: What does the indicator read at that moment?

Any attempt to consistently answer this query fails. How the failure is reflected in any attempted physical instantiation will depend on the details of the physics—in one instantiation the indicator device will burn out before the crucial moment, in another the indicator will continue to display but the display will not faithfully mirror the information sent from γ_1, etc. But independently of the details of the physics, we know in advance that the functional description of the device is not self-consistent. Does this knowledge constitute a general reductio of the possibility of using M–H spacetimes to create the functional equivalents of Plato machines? No, for the inconsistency here can be traced to the conditions imposed on one component of the π machine—the receiver–indicator—and such conditions are not imposed in mimicking Plato machines.

If the M–H analogue of the super π machine is to operate as intended, then the receiver–indicator must satisfy three demands: (a) the indicator has a definite state for all relevant values of its proper time τ, (b) the indicator is faithful in the sense that, if it receives an odd/even signal at τ, then it instantly adopts the corresponding odd/even indicator state, and (c) the indicator does not change its state except in response to a received signal in the sense that if τ_{ns} is a time at which no signal is received, then the indicator state at τ_{ns} is the limit of indicator states as τ approaches τ_{ns} from below. These demands are supposed to guarantee that at the crucial moment the indicator displays the parity of the "last digit" of π. That such a component is possible by itself leads to contradictions if it is assumed that the receiver–indicator device is subject to infinitely many alternating signals in a finite time. The limit required by (c) does not always exist, contradicting (a). I take the impossibility of such a component to be the lesson of forlorn attempts to construct an M–H analogue of the super π machines.

Denying the use of such functionally inconsistent devices will not affect

attempts to construct M–H analogues of Plato machines and to use them to gain new mathematical knowledge. The computer γ_1 is an infinity machine in the innocuous sense that it performs an infinite number of operations in an infinite amount of proper time. I see no grounds for thinking that such machines involve any conceptual difficulties unless they are required to compute a non-existent quantity. The uses to which I will put them make no such demand. Similarly, a conceptually non-problematic receiver–indicator device can be coupled to the computer through M–H spacetime relations in order to determine the truth values of mathematical conjectures. To flesh out the suggestion already made above, imagine, as in Pitowsky's example, that γ_1 is the world line of a computer which successively checks a conjecture of number theory for '1', for '2', etc. Since it has available to it an infinite amount of proper time the computer will in the fullness of time check the conjecture for all the integers. It is arranged that γ_1 sends a signal to γ_2 if and only if a counterexample is found. γ_2 is equipped with a receiver and an indicator device that is initially set to 'true' and which retains that state unless the receiver detects a signal, when the indicator shifts to 'false' and the receiver shuts off. By reading the display at the M–H point, γ_2 can learn whether or not the conjecture is true. Although I can give no formal proof of the consistency of this functional description, I see no basis for doubt. However, I will show below that attempts to physically instantiate this functional description run into various difficulties. But the difficulties have nothing to do with the paradoxes and conundrums of Thomson lamps and the like.

4.5 Characterization of Malament–Hogarth spacetimes

It was seen in section 4.3 that M–H spacetimes are not globally hyperbolic and thus violate Penrose's version of strong cosmic censorship. The converse is generally not true: some spacetimes that are not globally hyperbolic can fail to be M–H spacetimes. (Trivial example: Minkowski spacetime with a closed set of points removed does not contain a Cauchy surface but is not an M–H spacetime.) Some M–H spacetimes are acausal. Gödel spacetime is causally vicious in that for every point $p \in M$ $(=\mathbb{R}^4)$ there is a closed future-directed timelike curve through p (see chapter 6). In fact, for any $p \in M$, $I^-(p) = M$. Since Gödel spacetime contains timelike half-curves of infinite proper length, every point is an M–H point. I will not discuss such acausal spacetimes here. The reason is not because I think that the so-called paradoxes of time travel show that such spacetimes are physically impossible; indeed, I will argue just the opposite in chapter 6. But such paradoxes do raise a host of difficulties which, though interesting in themselves, only serve to obscure the issues about supertasks I wish to emphasize.

Therefore, in what follows I will restrict attention to causally well-behaved spacetimes. In particular, all of the spacetimes I will discuss are

stably causal, which entails the existence of a global time function (see chapter 6). I claim that among such spacetimes satisfying some subsidiary conditions to be announced, the M–H spacetimes are physically characterized by divergent blueshifts. The intuitive argument for this assertion is straightforward. During her lifetime, γ_1 measures an infinite number of vibrations of her source, each vibration taking the same amount of her proper time. γ_2 must agree that an infinite number of vibrations take place. But within a finite amount of his proper time, γ_2 receives an infinite number of light signals from γ_1, each announcing the completion of a vibration. For this to happen γ_2 must receive the signals in ever decreasing intervals of his proper time. Thus, γ_2 will perceive the frequency of γ_1's source to increase without bound. (This argument does not apply to acausal M–H spacetimes. The simplest example to think about is the cylindrical spacetime formed from two-dimensional Minkowski spacetime by identifying two points (x_1, t_1) and (x_2, t_2) just in case $x_1 = x_2$ and $t_1 = t_2$ modulo π. γ_2 can be chosen to be some finite timelike geodesic segment and γ_1 can be a timelike half-geodesic that spirals endlessly around the cylinder. The light signals from γ_1 may arrive at γ_2 all mixed up and not blueshifted.)

The main difficulty with this informal argument, as with all of the early literature on the redshift/blueshift effect (see Earman and Glymour 1980) is that the concept of frequency it employs refers to the rate of vibration of the source at γ_1 and to the rates at which γ_1 sends and γ_2 receives signals. But the effect actually measured by γ_2 depends on the frequency of the light signal (photon) itself. Thus, we need to calculate the blueshift using the definition of the emission frequency of a photon from a point $p_1 \in \gamma_1$ as $\omega_1 =: -(k_a V_1^a)|_{p_1}$ and the measured frequency of the photon as received at the point $p_2 \in \gamma_2$ as $\omega_2 =: -(k_a V_2^a)|_{p_2}$, where the timelike vectors V_1^a and V_2^a are respectively the normed tangent vectors to the world lines γ_1 and γ_2, and the null vector k^a is the tangent to the world line of the photon moving from the first to the second observer (see Fig. 4.2). Then the redshift/blueshift effect is given by the ratio

$$\frac{\omega_2}{\omega_1} = \frac{(k_a V_2^a)|_{p_2}}{(k_a V_1^a)|_{p_1}} \tag{4.1}$$

The following key fact is established in the appendix at the end of this chapter.

LEMMA 4.2. *Let M, g_{ab} be a Malament–Hogarth spacetime containing a timelike half-curve γ_1 and another timelike curve γ_2 from point q to point p such that $\int_{\gamma_1} d\tau = \infty$, $\int_{\gamma_2} d\tau < \infty$, and $\gamma_1 \subset I^-(p)$. Suppose that the family of null geodesics from γ_1 to γ_2 forms a two-dimensional integral submanifold in which the order of emission from γ_1 matches the order of reception at γ_2. If the photon frequency ω_1 as measured by the sender γ_1 is constant, then the time-integrated photon frequency $\int^{p_2} \omega_2 \, d\tau$ as measured by the receiver γ_2 diverges as p_2 approaches p.*

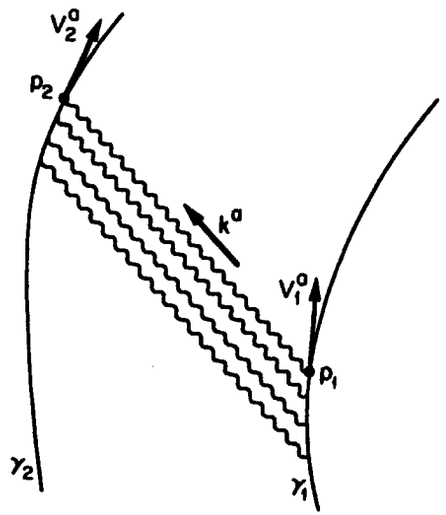

Fig. 4.2 The redshift/blueshift effect

Parameterize γ_2 by a t such that γ_2's past endpoint q corresponds to $t = 0$ and p corresponds to $t = 1$. Then it follows from Lemma 4.2 that $\lim_{t \to 1} \omega_2(t) = \infty$ if the limit exists. If not, then $\lim_{t \to 1} \omega^{\text{lub}}(t) = \infty$, where $\omega^{\text{lub}}(t) =: \text{lub}\{\omega_2(t') : 0 < t' \leq t\}$. Thus, one can choose on γ_2 a countable sequence of points approaching p such that the blueshift as measured by γ_2 at those points diverges. Typically this behavior will hold for any such sequence of points on γ_2, but there are some mathematically possible M–H spacetimes where γ_2 measures no red or blueshift at some sequence of points approaching p.

The following example (due to R. Geroch and D. Malament) illustrates this counterintuitive feature. As in the toy model in Fig. 4.1, start with parallel timelike geodesics of Minkowski spacetime. Parameterize γ_1 by the proper time τ of the Minkowski metric and adjust the curve so that the past endpoint corresponds to $\tau = 0$ and r corresponds to $\tau = 1$. At the points on γ_1 corresponding to $\tau = \tau_n = 1 - (3/4)(1/2^n)$ draw a sphere of radius $r_n = 1/2^{n+3}$ (as measured in the natural Euclidean metric). On the nth sphere put a conformal factor Ω_n which goes smoothly to 1 on the surface of the sphere and which has its maximum value at the point on γ_1 corresponding to τ_n. Construct the Ω_n such that the proper time along γ_1 in the conformal metric $\Omega^2 \eta_{ab}$ is infinite. For instance, if γ_1^n is the part of γ_1 within the nth sphere, set Ω_n so that $\int_{\gamma_1^n} \Omega_n \, d\tau = 1$. The result is an M–H spacetime. But at the points on γ_2 that receive photons from the points on γ_1 corresponding to $\tau = 1/2, 3/4, 7/8$, etc., there is no blue- or redshift.

While mathematically well defined, such examples are physically pathological. In particular, I do not know of any examples of M–H spacetimes which are solutions to Einstein's field equations for sources satisfying standard

energy conditions (see section 4.6) and which have the curious feature that the blueshift as measured by γ_2 diverges along some but not all sequences of points approaching the M–H point. Thus, although the slogan that M–H spacetimes involve divergent blueshifts is literally incorrect, it is essentially correct in spirit.

It may help to fix intuitions by computing the blueshift in some concrete examples. For the toy model pictured in Fig. 4.1 the result is

$$\frac{\omega_2}{\omega_1} = \frac{\Omega_{p_1}}{\Omega_{p_2}} = \Omega_{p_1} \tag{4.2}$$

This ratio diverges as γ_1 approaches the (missing point) r and γ_2 approaches the M–H point p.

Another stably causal M–H spacetime is obtained by taking the universal covering of anti-de Sitter spacetime (Hawking and Ellis 1973, pp. 131–134). Suppressing two spatial dimensions, the line element can be written as $ds^2 = dr^2 - (\cosh^2 r)\, dt^2$. Following Hogarth (1992) we can take γ_2 to be given by $r = r_2 =$ constant and γ_1 to be given by a solution to $dr/dt = \cosh r/\sqrt{2}$ (see Fig. 4.3). The blueshift is

$$\frac{\omega_2}{\omega_1} = \frac{\cosh r_1}{\cosh r_2(\sqrt{2} - 1)} \tag{4.3}$$

which diverges as $r_1 \to \infty$ and p_2 approaches the M–H point p.

We can also pose the converse question as to whether a divergent blueshift behavior indicates that the spacetime has the M–H property. The answer is positive in the sense that the proof of Lemma 4.2 can be inverted.

The fact that an M–H spacetime gives an indefinitely large blueshift for the photon frequency implies that the spacetime structure acts as an arbitrarily powerful energy amplifier. This might seem to guarantee

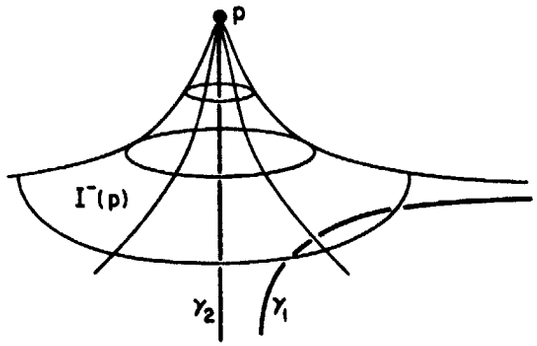

Fig. 4.3 Anti-de Sitter spacetime is a Malament–Hogarth spacetime

unambiguous communication from γ_1 to γ_2. But this first impression neglects the fact that a realistic instantiation of γ_1 will have thermal properties. The slightest amount of thermal radiation will be amplified indefinitely, which will tend to make communication impossible. In order not to destroy the receiver at γ_2, γ_1 will have to progressively reduce the energy of the photons she sends out. This means that there will be a point at which the energy of the signal photons will be reduced below that of the thermal noise photons. The indefinite amplification of the thermal noise will in any case destroy the receiver. Perhaps this difficulty can be met by cooling down γ_1 so as to eliminate thermal noise or by devising a scheme for draining off the energy of the signal photon while in transit. But even assuming a resolution of this difficulty, still further problems dog the attempt to use M–H spacetimes to accomplish supertasks.

4.6 Supertasks in Malament–Hogarth spacetimes

Are supertasks in Malament–Hogarth spacetimes to be taken seriously? The question involves three aspects. The first concerns whether M–H spacetimes are physically possible and physically realistic. As a necessary condition for physical possibility, general relativists will want to demand that the spacetime be part of a solution to Einstein's field equations for a stress–energy tensor T^{ab} satisfying some form of energy condition, weak, strong, or dominant (see chapter 3). The toy model of Fig. 4.1 can be regarded as a solution to Einstein's field equations with vanishing cosmological constant Λ by computing the Einstein tensor $G_{ab}(g)$ and then defining $T_{ab} =: (1/8\pi)G_{ab}$. But as conjectured in chapter 3, such models may be ruled out by the energy conditions. Anti-de Sitter spacetime, another M–H spacetime, can be regarded as a vacuum solution to Einstein's field equations with $\Lambda = R/4$, R (<0) being the curvature scalar; then the energy conditions are trivially satisfied. However, if it is required that $\Lambda = 0$, anti-de Sitter spacetime is ruled out by the strong energy condition if a perfect fluid source is assumed (see chapter 3).

None of these concerns touch Reissner–Nordström spacetime which is the unique spherically symmetric electrovac solution of Einstein's field equations with $\Lambda = 0$ (Hawking and Ellis 1973, pp. 156–161). Since this spacetime is an M–H spacetime, at least some M–H spacetimes meet the minimal requirements for physical possibility.

It is far from clear, however, that M–H spacetimes meet the (necessarily vaguer) criteria for physically realistic spacetime arenas. For one thing, it was seen in the preceding section that M–H spacetimes involve divergent blueshifts, which may be taken as an indicator that these spacetimes involve instabilities. Such is the case with Reissner–Nordström spacetime, where a small perturbation on an initial value hypersurface Σ (see Fig. 4.4) can produce an infinite effect on the future Cauchy horizon $H^+(\Sigma)$ of Σ (see Chandrasekhar and Hartle 1982).

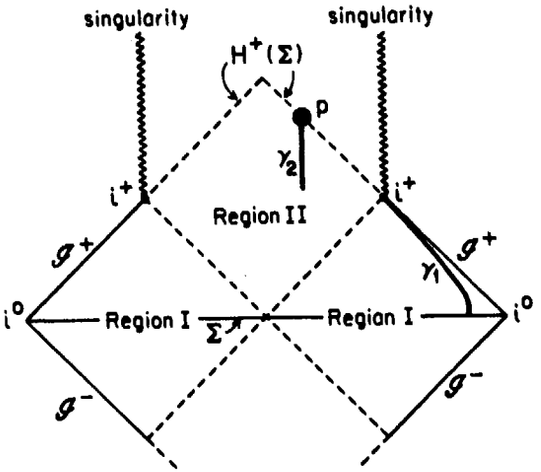

Fig. 4.4 Reissner–Nordström spacetime is a Malament–Hogarth spacetime

For another thing, various M–H spacetimes run afoul of one or other versions of Penrose's cosmic censorship hypothesis. By Lemma 4.1 all M–H spacetimes violate strong cosmic censorship, and many examples of M–H spacetimes violate weaker versions as well. Thus, evidence that cosmic censorship holds for physically reasonable spacetimes is ipso facto evidence against the physical reasonableness of M–H spacetimes. Conversely, one might take the rather bizarre scenarios that can be concocted in M–H spacetimes as grounds for thinking that a cosmic censor should be at work. But then again, those with a taste for the bizarre may hope that cosmic censorship fails just so that they can own the functional equivalent of a Plato machine.

I now turn to the second aspect of the question of how seriously to take the possibility of completing supertasks in M–H spacetimes. This aspect concerns whether γ_1 can be implemented by a physically possible/physically realistic device which, over the infinite proper time available to it, carries out the assigned infinite task. Once again the question is made difficult by the fact that there is no agreed upon list of criteria that identify physically realistic devices. I will make the task of tackling this question tractable by confining attention to dynamical constraints that physically realistic γ_1 should satisfy. (One doesn't have to worry about dynamical constraints on γ_2 since typically γ_2 can be chosen to be a geodesic.) Minimally, the magnitude of acceleration of γ_1 must remain bounded, otherwise any device that one could hope to build would be crushed by g-forces. This condition is satisfied in the anti-de Sitter case (Fig. 4.3) where $a(r_1) = \sqrt{2}\,[\exp(2r_1) - 1]/[\exp(2r_1) + 1]$, which approaches $\sqrt{2}$ as $r_1 \to \infty$. However, we must also demand a finite bound on the total acceleration of γ_1: $\mathrm{TA}(\gamma_1) =: \int_{\gamma_1} a\, d\tau$. For even with perfectly efficient rocket engines, the final mass m_f of the rocket and the mass m_{fuel} of

the fuel needed to accelerate the rocket must satisfy (see the appendix to this chapter)

$$\frac{m_f}{m_f + m_{fuel}} \leq \exp(-\mathrm{TA}(\gamma_1)) \tag{4.4}$$

Thus if $\mathrm{TA}(\gamma_1) = \infty$ an infinite amount of fuel is needed for any finite payload. In the anti-de Sitter case, $d\tau_1 = d\tau$ so that $\mathrm{TA}(\gamma_1) = \infty$, and the demand fails. In the toy model of Fig. 4.1 the demand is met since $\mathrm{TA}(\gamma_1) = 0$, γ_1 being a geodesic; but the spacetime involved was ruled out as not physically possible. In Reissner–Nordström spacetime a timelike geodesic γ_1 can be chosen to start on the time slice Σ (see Fig. 4.4) and to go out to future timelike infinity i^+.[5] And $\gamma_1 \subset I^+(p)$ for an appropriate point $p \in H^+(\Sigma)$. But again there are reasons to regard this spacetime as not being physically realistic.

Finally, since a physically realistic device must have some finite spatial extent, we are really concerned not with a single world line γ_1 but with a congruence Γ_1 of world lines. Even if Γ_1 is a geodesic congruence it cannot be instantiated by a physically realistic computer (say) unless the tidal forces it experiences remain bounded. Since the tidal forces are proportional to the Riemann curvature tensor,[6] one can satisfy this demand in Reissner–Nordström spacetime, which is asymptotically flat in the relevant region. One simply starts the geodesic congruence sufficiently far out towards spatial infinity and has it terminate on future timelike infinity i^+.

To summarize the discussion up to this point, it is not clear that any M–H spacetime qualifies as physically possible and physically realistic. But to the extent that M–H spacetimes do clear this first hurdle, it seems that the role of γ_1 can be played by a world line or world tube satisfying realistic dynamical constraints. However, Pitowsky (1990) feels that, for other reasons, γ_1 cannot be instantiated by a computer that will carry out the assigned infinite task. I will take up his worry in section 4.8 below. Before doing so I turn to the third aspect of whether M–H can be taken seriously. It concerns discriminations that the receiver γ_2 must make.

4.7 Malament–Hogarth spacetimes and unresolved mathematical conjectures

Can Malament–Hogarth spacetimes be used to gain knowledge of the truth values of unresolved mathematical conjectures? Suppose now for sake of discussion that some M–H spacetimes are regarded as physically possible and physically realistic and that in these arenas there are no barriers to a physically possible and physically realistic instantiation of γ_1 by a computer which carries out the task of checking Fermat's last theorem or some other unresolved conjecture of number theory. Nevertheless there are reasons to doubt that γ_2

can use γ_1's work to gain genuine knowledge of the truth value of the conjecture. The pessimism is based on a strengthening of Lemma 4.1.

LEMMA 4.3. *Suppose that $p \in M$ is a M–H point of the spacetime M, g_{ab} (that is, there is a future-directed timelike half-curve $\gamma_1 \subset M$ such that $\int_{\gamma_1} d\tau = \infty$ and $\gamma_1 \subset I^-(p)$). Choose any connected spacelike hypersurface $\Sigma \subset M$ such that $\gamma_1 \subset I^+(\Sigma)$. Then p is on or beyond $H^+(\Sigma)$.*

PROOF: If $p \in \mathrm{int}[D^+(\Sigma)]$ then there is a $q \in D^+(\Sigma)$ which is chronologically preceded by p. $M' =: (I^-(q) \cap I^+(\Sigma)) \subset D^+(\Sigma)$, and the smaller spacetime $M', g_{ab}|_{M'}$ is globally hyperbolic. Choose a Cauchy surface Σ' for this smaller spacetime which passes through p. Since $\gamma_1 \subset M'$ we can proceed as in the proof of Lemma 1 to obtain a contradiction.

Lemma 4.3 is illustrated by the Reissner–Nordström spacetime (Fig. 4.4). Any M–H point involved with a γ_1 starting in region I must lie on or beyond $H^+(\Sigma)$.

Think of Σ as an initial value hypersurface on which one specifies initial data that, along with the laws of physics, prescribes how the computer γ_1 is to calculate and how it is to signal its results to γ_2. Since by Lemma 4.3 any M–H point $p \in \gamma_2$ must lie on or beyond $H^+(\Sigma)$ for any appropriate Σ, events at p or at points arbitrarily close to p are subject to non-deterministic influences. In typical cases such as the Reissner–Nordström spacetime illustrated in Fig. 4.4 there are null rays which pass arbitrarily close to any $p \in H^+(\Sigma)$ and which terminate in the past direction on the singularity. There is nothing in the known laws of physics to prevent a false signal from emerging from the singularity and conveying the misinformation to γ_2 that a counter-example to Fermat's conjecture has been found.[7] (γ_2 need not measure an infinite blueshift for photons emerging from the singularity; at least there is nothing in Lemma 4.2 or the known laws of physics that entails such a divergent blueshift.) Of course, the receiver γ_2 can ignore the signal if he knows that it comes from the singularity rather than from γ_1. But to be able to discriminate such a false signal from every possible true signal that might come from γ_1, γ_2 must be able to make arbitrarily precise discriminations. In the original situation it was the Plato machine that had to perform a supertask by compressing an infinite computation into a finite time span. The trick adopted here was to finesse the problems associated with such a supertask by utilizing two observers in relativistic spacetime. But we have found that the finesse also involves a kind of supertask—not on the part of the computer but on the part of the receiver who tries to use the work of the computer to gain new mathematical knowledge.

This verdict may seem unduly harsh. If γ_2 is to be sure beforehand that, whatever γ_1's search procedure turns up, he will obtain knowledge of the truth value of Fermat's conjecture, then γ_2 must be capable of arbitrarily precise discriminations. But, it may be urged, if γ_2 is capable of only a finite

degree of precision in his signal discriminations, he may yet learn that Fermat's conjecture is false (if indeed it is) if he receives a signal long enough before the M–H point so that it lies within his discrimination range. This, however, would be a matter of good fortune. One can pick at random a quadruple of numbers (x, y, z, n), $n \geq 3$, and check whether $x^n + y^n = z^n$. If one is lucky, a counterexample to Fermat's conjecture will have been found. But the interest in Platonist computers and their M–H analogues lay in the notion that they do not rely on luck.

Of course any observer faces the problem of filtering out spurious backgound signals from those genuinely sent from the system observed. But it is usually assumed that sufficiently thorough attention to the experimental setup could at least in principle control all such signals. What Lemma 4.3 shows, however, is that no such efforts can succeed even in principle in our case. No matter how carefully and expansively we set up our experiment—hat is, no matter how large we choose our initial value hypersurface—we cannot prevent spurious signals from reaching p or coming arbitrarily close to p.

The problem can be met by means of a somewhat more complicated arrangement between γ_1 and γ_2 by which γ_1 not only sends a signal to γ_2 to announce the finding of a counterexample but also encodes the quadruple of numbers that constitutes the counterexample. A false signal may emerge from the singularity, but γ_2 can discover the falsity by a mechanical check. With the new arrangement γ_2 no longer has to discriminate where the signal came from since a counterexample is a counterexample whatever its origin. Unfortunately, γ_2 may still have to make arbitrarily fine discriminations since the quadruple sent will be of arbitrarily great size (=number of bits) and must be compressed into a correspondingly small time interval at γ_2.

The worry about whether γ_2 can gain knowledge of Fermat's conjecture by using γ_1's efforts also involves the concern about γ_2's right to move from 'γ_1 has not sent me a signal' to 'Fermat's conjecture is true'. The correctness of the inference is not secured by the agreement γ_1 and γ_2 have worked out, for even with the best will in the world γ_1 cannot carry out her part of the agreement if events conspire against her. As suggested above, the most straightforward way to underwrite the correctness of the inference is for there to be a spacelike Σ such that $\gamma_1 \subset D^+(\Sigma)$ and such that initial conditions on Σ together with the relevant laws of physics guarantee that γ_1 carries out her search task. And if, as is compatible with at least some M–H spacetimes (e.g., Reissner–Nordström spacetime), the M–H point p can be chosen so that $\Sigma \subset I^-(p)$, it would seem that γ_2 could in principle come to know that the conditions which underwrite the inference do in fact obtain. But the rub is that p or points arbitrarily close to p may receive a false signal from the singularity, indicating that conditions are not conducive to γ_1's carrying out her task. If so, γ_2 is not justified in making the inference unless he can discriminate false signals as such. This, of course, is just another version of the difficulty already discussed. But the present form does not seem to have an easy resolution.

4.8 Can γ_1 carry out the assigned task?

γ_1 is supposed to check an unresolved conjecture of number theory for each of the integers. By construction, γ_1 has time enough. But Pitowsky feels that γ_1 never has world enough.

> The real reason why Platonist computers are physically impossible *even in theory* has to do with the computation space. According to general relativity the material universe is finite. Even if we use the state of every single elementary particle in the universe, to code a digit of a natural number, we shall very soon run out of hardware. (Pitowsky 1990, p. 84)

In response, I note that general relativity theory does not by itself imply a spatially or materially finite universe. Further, it was seen that there are spatially infinite M–H spacetimes, such as Reissner–Nordström spacetime, that are live physical possibilities in the minimal sense that they satisfy Einstein's field equations and the energy conditions. A γ_1 who wanders off into the asymptotically flat region of this spacetime certainly has space enough for any amount of hardware she needs to use. But she cannot avail herself of an unlimited amount of hardware without violating the implicit assumption of all of the foregoing; namely, that γ_1 and γ_2 have masses so small that they do not significantly perturb the background metric. Here Pitowsky's objection has some bite.

Perhaps there are solutions to Einstein's field equations where the spacetime has the M–H property and there is both space enough and material enough for a physically embodied computer with an unlimited amount of computation space.[8] Pending the exhibition of such models, however, one must confine oneself to tasks that can be accomplished in an infinite amount of time but with a finite amount of computation space. Whether there are such tasks that deserve the appellation 'super' remains to be seen.

4.9 Conclusion

Thomson lamps, super π machines, and Platonist computers are playthings of philosophers; they are able to survive only in the hothouse atmosphere of philosophy journals. In the end, M–H spacetimes and the supertasks they underwrite may similarly prove to be recreational fictions for general relativists with nothing better to do. But in order to arrive at this latter position requires that one first resolve some of the deepest foundation problems in classical general relativity, including the nature of singularities and the fate of cosmic censorship. It is this connection to real problems in physics that makes them worthy of discussion.

There are also connections to the philosophy of mathematics and to the theory of computability. Because of finitist scruples, some philosophers have doubted that it is meaningful to assign a truth value to a formula of

arithmetic of the form $(\exists x_1)(\exists x_2)\ldots(\exists x_n)F(x_1, x_2, \ldots, x_n)$. It seems to me unattractive to make the truth of mathematical statements depend on the contingencies of spacetime structure. The sorts of arrangements considered above can be used to decide the truth value of assertions of arithmetic with a prenex normal form that is purely existential or purely universal.[9] (Fermat's last theorem, for example, has a purely universal form.) For such an assertion γ_1 is set to work to check through the (countably infinite) list of n-tuples of numbers in search of a falsifier or a verifier according as the assertion to be tested is universal or existential, and γ_1 reaps from these labors a knowledge of the truth value of the assertion. But as soon as mixed quantifiers are involved, the method fails.[10] However, Hogarth (1994) has shown how more complicated arrangements in general relativistic spacetimes can in principle be used to check the truth value of any arithmetic assertion of arbitrary quantificational complexity. Within such a spacetime it is hard to see how to maintain the attitude that we do not have a clear notion of truth in arithmetic.[11]

The computational arrangements between γ_1 and γ_2 envisioned might also seem to bring into doubt Church's proposal that effective/mechanical computability is to be equated with Turing computability or recursiveness, for apparently γ_1 and γ_2 can in concert obtain a resolution to recursively unsolvable problems by means that certainly seem to merit the appellations of 'effective' and 'mechanical'. But putting the matter this way is a little unfair to Church since any account of effective/mechanical computability that implies that there are subsets of numbers which can be effectively/mechanically enumerated, but whose complements cannot be, will be subject to the one-upmanship of bifurcated supertasks. Perhaps the most illuminating way to state the moral to be drawn from bifurcated supertasks is that two levels of computation need to be distinguished: the first corresponding to what the slave computer γ_1 can do, the second to what γ_2 can infer by having causal access to all of γ_1's labors. Church's proposal is best construed as aimed at the first level and as asserting that Turing computability is an upper bound on what any physical instantiation of γ_1 can accomplish. Read in this way, there is nothing in present concerns to raise doubts about Church's proposal.[12]

Appendix: Proofs of Lemma 4.2 and Equation 4.4

Proof of Lemma 4.2. The null geodesics from γ_1 to γ_2 form a two-dimensional submanifold. For each of the null geodesics select an affine parameter λ which varies from 0 at γ_1 to 1 at γ_2. (This will always be possible since an affine parameter can be rescaled by an arbitrary linear transformation.) The null propagation vector $k^a = (\partial/\partial\lambda)^a$ satisfies the geodesic equation

$$k^a\nabla_a k^b = 0 \tag{A4.1}$$

By supposition, these null geodesics form a submanifold. By connecting points of equal λ values, form a family of curves indexed by λ that covers the submanifold and interpolates between γ_1 and γ_2. Select any parameterization t of γ_1 and propagate this parameterization along the null geodesics to all the interpolating curves so that each null geodesic passes through points of equal t value. The indices λ and t form a coordinate system for the two-manifold. k^a and $\zeta^a = (\partial/\partial t)^a$ are its coordinate basis vector fields, which entails that they satisfy the condition $[\zeta, k]^a = 0$ so that

$$\zeta^a\nabla_a k_b - k^a\nabla_a\zeta_b = 0 \tag{A4.2}$$

It follows that $(\zeta_a k^a)$ is a constant along the photon world lines. To show this it needs to be demonstrated that

$$\frac{d}{d\lambda}(\zeta_a k^a) = k^a\nabla_a(\zeta_b k^b) = 0 \tag{A4.3}$$

This is done by computing

$$k^a\nabla_a(\zeta_b k^b) = k^b(k^a\nabla_a\zeta_b) + \zeta_b k^a\nabla_a k^b \tag{A4.4}$$

The second term on the right-hand side of (A4.4) vanishes in virtue of (A4.1). Equation (A4.2) can then be used to rewrite the first term on the right-hand side as $\zeta^a k^b\nabla_a k_b = \frac{1}{2}\zeta^a\nabla_a(k_b k^b) = 0$ since k^a is a null vector.

Thus, for a photon sent from γ_1 to γ_2, we have $k_a\zeta_1^a = k_a\zeta_2^a$, or $k_a V^a|\zeta_2^a| = k_a V^a|\zeta_2^a|$, where $V^a =: \zeta^a/|\zeta^a|$ is the normed tangent vector to the timelike world line. So from the definition (4.1) of photon frequency ratios one can conclude that $\omega_1|\zeta_1^a| = \omega_2|\zeta_2^a|$ which implies that

$$\int_{\gamma_1}\omega_1|\zeta_1^a|\,dt = \int_{\gamma_2}\omega_2|\zeta_2^a|\,dt \tag{A4.5}$$

or

$$\int_{\gamma_1}\omega_1\,d\tau = \int_{\gamma_2}\omega_2\,d\tau \tag{A4.6}$$

But $\int_{\gamma_1} d\tau = \infty$ and $\int_{\gamma_2} d\tau < \infty$. So if ω_1 is constant along γ_1, (A4.6) can hold only if $\int\omega_2\,d\tau = \infty$.

Proof of (4.4) (from Malament 1985). If V^n is the (normalized) four-velocity of the rocket and m its mass, the rate at which its energy–momentum changes is $V^p\nabla_p(mV^n)$, which must balance the energy–momentum $\mathcal{J}^n$ of its exhaust (it being assumed that the rocket's motor is the only source of

propulsion). Thus,

$$\mathcal{J}^n = V^n(V^p\nabla_p m) + mA^n \tag{A4.7}$$

where $A^n = V^m\nabla_m V^n$ is the four-acceleration. Since $\mathcal{J}^n$ is not spacelike, $\mathcal{J}^n\mathcal{J}_n \leq 0$. Consequently,

$$-(V^p\nabla_p m)^2 + m^2a^2 \leq 0 \tag{A4.8}$$

which uses $V^nV_n = -1$, $V^nA_n = 0$, and $a =: (A^nA_n)^{1/2}$. Furthermore, because the rocket is using up fuel, $V^p\nabla_p m \leq 0$. Thus,

$$a \leq -V^p\nabla_p(\ln(m)) = -\frac{d}{d\tau}(\ln(m)) \tag{A4.9}$$

So if m_i and m_f are the initial and final masses of the rocket, integration of (A4.9) yields

$$\mathrm{TA}(\gamma) \leq \ln(m_i/m_f) \tag{A4.10}$$

And since $m_i = m_f + m_{fuel}$,

$$\frac{m_f}{m_f + m_{fuel}} \leq \exp(-\mathrm{TA}(\gamma)) \tag{A4.11}$$

Notes

1. This task is taken up in Earman and Norton (1994).

2. A few new wrinkles are added in Earman and Norton (1994) concerning Ross's paradox (see Allis and Koetsier 1991; van Bendegem 1994) and some other paradoxes of the infinite.

3. A subset of $S \subset \mathbb{N}$ is said to be *recursively enumerable* (r.e.) just in case it is the range of a (partial) recursive function $f: \mathbb{N} \to \mathbb{N}$; informally, this means that there is an effective procedure for generating the members of S. S is said to be *recursive* if both S and the complement of S are r.e.; informally, there is an effective procedure for deciding membership in S. Key results on undecidability of formal systems hinge on there being sets that are r.e. but not recursive.

4. David Malament has pointed out that a quick proof of Lemma 4.1 can be obtained by using Prop. 6.7.1 of Hawking and Ellis (1973): For a globally hyperbolic spacetime, if $p \in \mathcal{J}^+(q)$, then there is a non-spacelike geodesic from q to p whose length is greater than or equal to that of any other non-spacelike curve from q to p. Suppose that $\gamma \in I^-(p)$ is a timelike half-curve with endpoint q and that $\int_\gamma d\tau = \infty$. Since the endpoint q of γ belongs to $I^-(p)$, we could apply the proposition to p and q if the spacetime were globally hyperbolic. But then a contradiction results, since whatever the bound on the length of the timelike geodesic from q to p, we could exceed it by going along γ sufficiently far and then over to p. Robert Wald noted

that an even quicker proof is obtained from the compactness of $\mathcal{J}^-(p) \cap \mathcal{J}^+(q)$ together with strong causality, which are consequences of global hyperbolicity.

5. In Fig. 4.4 i^0 labels spatial infinity and $\mathscr{I}^+$ and $\mathscr{I}^-$ respectively label future and past null infinity.

6. See Wald (1984, pp. 46–47) for a derivation of the formula for geodesic deviation.

7. One might also worry that a burst of noise from the singularity could swamp an authentic signal. But since any real signal arrives at γ_2 prior to the singularity noise, the former is not masked by the latter as long as the receiver can discriminate between a signal and noise.

8. The considerations raised here are similar to those discussed by Barrow and Tipler (1986) under the heading of "omega points."

9. Assuming that the relation quantified over is effectively decidable.

10. Showing this requires a more careful specification of how bifurcated infinity machines operate; see Earman and Norton (1994).

11. See Earman and Norton (1994) for more discussion of this and related matters.

12. But there are independent reasons to doubt Church's proposal; see Earman (1986) and Pitowsky (1990).

5

The Big Bang and the Horizon Problem

5.1 Introduction

Discussions from the recent astrophysics literature on observability, horizons, and the like have trickled down to the level of philosophical consciousness, at least in a folkloric way. For example, the folklore contains the wisdom that the standard big bang cosmological models contain particle horizons, that when this fact is coupled with the observed isotropy of the 3 K cosmic background radiation there arises a "horizon problem," and that inflationary cosmology solves this problem. Like most folklore, this example contains some truth and a number of distortions. There is a need to set the record straight and to correct some fundamental misimpressions about observability and horizons in relativistic cosmological theories. But more is at stake than correcting some misimpressions. The horizon problem provides an interesting test case for accounts of scientific explanation, for the perception of a "problem" in connection with particle horizons in standard big bang cosmology depends on views about what features a good scientific explanation should have. While this matter has received little attention in the philosophical literature, it has played an important role in guiding research in relativistic cosmology. The horizon problem is also a good test case for the principle of common cause since cosmological models with particle horizons provide examples where there is no common causal past for correlated events.

Sections 5.2 through 5.4 give a preliminary treatment of observability and horizons in relativistic cosmological models and introduce the concepts needed to assess the horizon problem. That problem is introduced in section 5.5 and then elaborated and diagnosed in sections 5.6 through 5.9. Section 5.10 reviews various strategies that astrophysicists have adopted in attempts to reach a resolution. Section 5.11 treats in detail the currently favored approach, inflationary cosmology. Section 5.12 confronts the issue of whether or not inflationary scenarios resolve the horizon problem. Section 5.13 contains conclusions and closing remarks.

5.2 Observability and light cones

We receive most of our information about the large-scale structure of spacetime on a series of light cones. How much does the information which, in principle, we can gather on these light cones allow us to infer about the structure of space and time? Remarkably, GTR yields the answer that the light cone data in conjunction with EFE allow us in principle to determine the spacetime geometry on and inside our past light cone (see Kristian and Sachs 1966; Ellis, Maartens, and Nel 1978).

Folkloric understanding of relativity theory says that this is all that direct observation and the laws of physics allow us to know about. As usual, idealize an observer as a timelike world line γ. Then two folklore dogmas can be stated. The first dogma is that at $p \in \gamma$, an observer whose world line is γ cannot have direct observational knowledge of events at a spacetime location $q \notin J^-(p)$.[1] This dogma seems to be firmly rooted in good common sense; for whatever else observation means, it must involve a causal link between observer and the events observed, and the fact that $q \notin J^-(p)$ means that q cannot be joined to p by a causal signal. (The possibility of faster-than-light particles is ignored here.) The second dogma is that if secure knowledge is limited to what can be deduced from direct observational knowledge plus the laws of physics, then secure knowledge at p does not extend to events at $q \notin J^-(p)$ because the state at q is not determined by the laws on the basis of the state in $J^-(p)$ and because, more generally, relativistic laws are inconsistent with action at a distance and, thus, do not constrain events happening in relatively spacelike regions of spacetime.

These folklore dogmas are, with some caveats and qualifications, basically correct. Seeing where the caveats and qualifications are needed helps to prepare the discussion of the horizon problem. The more sophisticated folklore takes account of an exception to the second dogma that arises when $J^-(p)$ contains a Cauchy surface Σ. In that case the observed state on Σ plus the coupled Einstein–matter equations will determine the state throughout the spacetime.[2] A necessary but not sufficient condition for such a case is that the universe be spatially finite in the sense that Σ is compact; and, since Σ is a Cauchy surface, the spacetime manifold M must be diffeomorphically $\Sigma \times \mathbb{R}$ (Dieckmann 1988) as illustrated by Fig. 5.1. The spacetime in this illustration is created by artifice—specifically, by identifying points in a larger spacetime. One would expect that examples of spacetimes which contain a (necessarily) compact Cauchy surface contained in the causal past of a point and which are not created by this artifice are relatively rare among the solutions to Einstein's field equations.

The second dogma is mistaken in a more fundamental way: paradigm examples of relativistic laws, such as Maxwell's equations of electromagnetism and EFE for gravitation, do constrain relatively spacelike events. In both cases the full set of equations can be divided into ***constraint equations***, which impose conditions on instantaneous initial data, and ***dynamical equations*** that

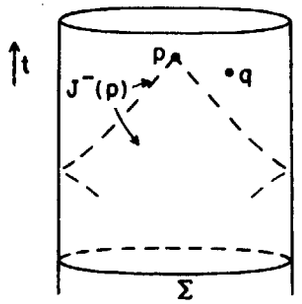

Fig. 5.1 At p sure prediction of events at q is possible

govern the temporal evolution of the data.[3] And in both cases the dynamical equations entail that if the constraint equations are satisfied for one instant, then they are satisfied at all instants. The simplest illustration of how the constraint equations work concerns electromagnetism in the special theory of relativity (STR). The constraint equations are $\nabla \cdot \mathbf{E} = \rho$ and $\nabla \cdot \mathbf{B} = 0$, where $\mathbf{E}$ and $\mathbf{B}$ are respectively the electric and magnetic field strengths and ρ is the electric charge density. Referring to Fig. 5.2, imagine that in the slice

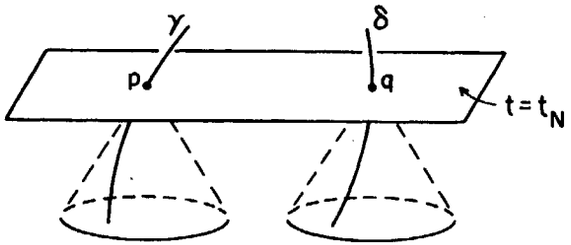

Fig. 5.2 Charged particles interacting in Minkowski spacetime

$t = t_N$ (N for "now") of Minkowski spacetime a sphere centered on q is drawn. (Here t is the time coordinate from some inertial coordinate system.) It follows from the first constraint equation that the electric flux through the sphere must be equal to the charge inside the sphere—in this case the charge on the particle whose world line is δ. Conversely, knowledge of $\mathbf{E}$ at the spacetime points lying on the sphere allows one to infer the charge contained inside, illustrating how relatively spacelike electromagnetic events are mutually constrained by the laws of electromagnetism.

A modification of this example also serves to challenge the first dogma of observation. Create a past-truncated Minkowski spacetime by deleting all those spacetime points r such that $t(r) \leq 2000$ B.C. (see Fig. 5.3). Suppose that γ and δ are the only particles in the universe and that both are electrically charged. Then although at p the observer whose world line is γ cannot *see* δ

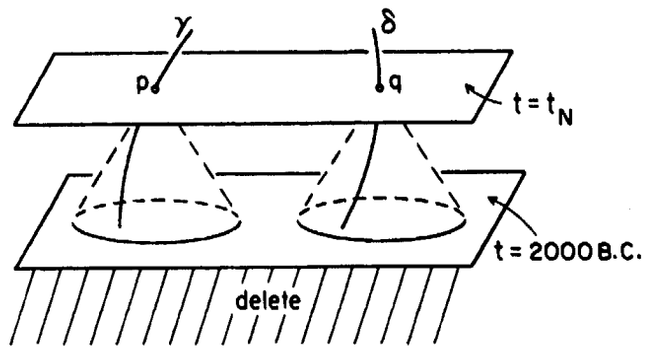

Fig. 5.3 Charged particles interacting in past-truncated Minkowski spacetime

since $\mathcal{J}^-(p) \cap \delta = \varnothing$, this observer can *feel* δ (see Ellis and Sciama 1972). For by the above argument, the Coulomb field of δ will have to be non-zero at p (or at any rate at a set of points of positive measure on any closed surface in $t = t_N$ containing q, so p might just as well be one of these points).

The first dogma can be defended in two ways against the notion that this feeling is observing. It can be reiterated that observation requires causal connection, and then it can be asserted that since by construction $\mathcal{J}^-(p) \cap \delta = \varnothing$, the causal connection is missing. But this defense begs the question at issue. Granted that in the present example there is no causal link forged by a causal signal, there certainly is a lawlike connection between the unseen δ and the tug felt by γ at p. Trying to decide whether this lawlike connection amounts to a causal connection by appealing to existing philosophical analyses of causation does not look promising. For instance, Lewis' (1975) counterfactual analysis of causation would have us try to decide the matter by asking (roughly) whether the nearest possible world to that of Fig. 5.3 in which there is no charge at q is also a world in which γ feels no tug at p. My intuition is that the answer is yes. But I do not put much store in such intuitions, and if issues about observability really turn on such subjective and context dependent notions as nearness or similarity of possible worlds, then arguing over them does not seem worth the candle.

A second and better defense of the first dogma would start with the position that genuine observational knowledge requires not only true belief but also that the belief has been formed by a reliable process. But just from the electric force felt by γ at p, γ cannot reliably infer the existence of a charged particle outside of $\mathcal{J}^-(p)$; for the force might be due to source-free electromagnetic radiation (which in any case must be present in the example of Fig. 5.3, as will be discussed in section 5.8). By exploring more of the electric field, γ can make a more reliable inference to the existence of an unseen charge; but by the time the exploration is extensive enough to make the inference completely secure, signals from δ will have had time to reach γ.

This defense fails in some spacetimes. Consider a modification of the

Minkowski metric in which the line element becomes $ds^2 = \Psi(x, y, z) \times (dx^2 + dy^2 + dz^2) - dt^2$.[4] If the scale factor $\Psi(x, y, z)$ is chosen appropriately, the sphere through q can have a small area even though p is a large distance from q. Then before a signal from δ can reach him, our observer may be able to sample the electric field on the entire sphere and, thus, can obtain certain knowledge of the existence of a charge within the sphere. However, because of the special nature of this example, the proponents of the standard dogmas may not be much moved. For most purposes then, I think we can agree with the folklore dogma that the limits of the observable at p are set by $\mathcal{J}^-(p)$.

The past-truncated Minkowski model of Fig. 5.3 is, of course, very artificial (at least for those who do not subscribe to the creationist line of Protestant fundamentalism). But it does serve to illustrate some general features of the particle horizons that occur naturally in general relativistic cosmological models. And in the electromagnetic example the astute reader can already detect the seeds of a problem that will be discussed in detail below in later sections.

5.3 What can we predict about the future?

For present purposes take prediction to mean deterministic prediction from the laws of physics. This implies that in a spacetime M, g_{ab} if events at $p \in M$ are to be predicted from the state in a region $X \subset M$, then it is necessary that p belong to the future domain of dependence $D^+(X)$ of X; for if $p \notin D^+(X)$ there will be possible causal influences that could effect the state at p without registering on X. If Σ is a Cauchy surface for M, g_{ab} then $D^+(\Sigma)$ includes every point in the spacetime to the future of Σ. So we may now put the question: In the deterministic sense of prediction, what can an observer γ predict at some point $p \in \gamma$ from her knowledge of events in $\mathcal{J}^-(p)$? In the special case where there is a Cauchy surface $\Sigma \subset \mathcal{J}^-(p)$ the observer is in a position to give a deterministic prediction of the entire future of the universe. But in general the answer is: Nothing![5] The formal point is that in Minkowski spacetime and in typical general relativistic spacetimes, $D^+(\mathcal{J}^-(p)) = \mathcal{J}^-(p)$ for any spacetime point p so that the only events which can be predicted from p are those that, from the perspective of p, have already happened—which is to say that the prediction is not a genuine prediction. A concrete example to illustrate the point for Minkowski spacetime is given in Fig. 5.4. In section 5.2 we saw that it is not true that there can be nothing in $\mathcal{J}^-(p)$ to alert γ to the presence of the photon α and the massive particle β whose world lines are past endless but never enter $\mathcal{J}^-(p)$. But the information furnished by typical relativistic constraint equations is weak, and in general the information will not be strong enough to tell γ for sure whether a photon or a massive particle will intersect her world line at some chosen point to the future of p.[6]

What then is the status of the forecasts we routinely make about the future? While the success of our predictions is due to more than lucky guessing,

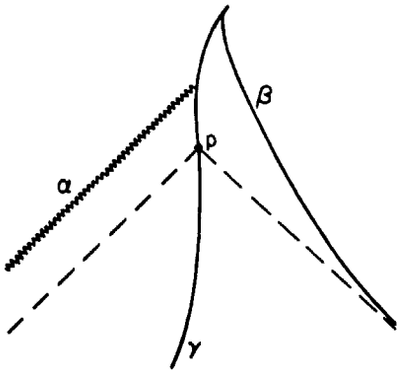

Fig. 5.4 γ's prediction at p of her future is foiled by the photon α and the massive particle β

there is also a large element of wishful thinking that gets dignified by such big names as the Copernican principle. According to this principle, we do not occupy a privileged position in the universe. Thus, if no world lines of photons or massive particles enter $\mathcal{J}^-(p)$, we may count ourselves as reasonable in concluding at p that no such particles exist outside of $\mathcal{J}^-(p)$ and that we will not be rudely surprised as is the observer in Fig. 5.4 who relies on this principle at p. Put in the negative mode, this principle is appealing—it modestly denies that we have some special status. But this seeming modesty is belied by the immodest use to which the principle was put in justifying an inductive extrapolation.

Can we do better than appeal to big name principles? Perhaps we can have reasons to think that there are no source-free photons like α in Fig. 5.4. If so, we are on the road to a more secure forecast. For the only way that a (past endless) world line of a massive source of photons could fail to fall into $\mathcal{J}^-(p)$ is for it to behave like β in Fig. 5.4. But the total integrated acceleration, from $t = -\infty$ to the present, of a world line like β is infinite. Thus, an infinite amount of energy would be needed to produce such a trajectory; or, if β were the world line of a rocket ship, an infinite fuel-to-payload ratio would be required. This makes promising the prospect of proving results to the effect that prediction will not be undermined by well-understood physical mechanisms. For example, one could try to prove that a trajectory like that of β cannot result from electromagnetic forces produced by a system of charged particles, at least not in a way that does not leave enough tracks on $\mathcal{J}^-(p)$ to infer the existence of β.

The example illustrated in Fig. 5.4 is based on Minkowski spacetime. Things are much worse in a spacetime like the past-truncated spacetime of Fig. 5.3, where there are past endless timelike geodesics that do not enter $\mathcal{J}^-(p)$. Similarly, in general relativistic spacetimes which are not artificially past-truncated but which have particle horizons, prediction is much harder.

This is one reason to dislike particle horizons. But before the complaint can be explored it is necessary to define particle horizons.

5.4 Event and particle horizons

The *locus classicus* of the modern treatment of horizons in cosmology is Rindler (1956). According to Rindler's definition, a *future event horizon* (FEH(γ)) for an observer with world line γ is a hypersurface in spacetime "which divides all events into two non-empty classes: those which have been, are, or will be observable by [γ], and those that are forever outside [γ's] possible powers of observation" (p. 663). Leaving aside the caveats about "observation" discussed in section 5.2, we can take FEH(γ) to be given by the boundary between those events lying inside of $\mathcal{J}^-(\gamma)$ and those lying outside. Some examples will help to illustrate the definition.

Example 1. If γ has a future endpoint e, then one would like to think of FEH(γ) as the future boundary of $\mathcal{J}^-(e)$, i.e., $\overline{\mathcal{J}^-(e)} - I^-(e)$. When $\mathcal{J}^-(p)$, $p \in M$, is closed, this boundary will coincide with the *past light cone* $L^-(p)$ of p, i.e., the locus of the past-directed null geodesics through p.[7] However, it can happen that $\mathcal{J}^-(p)$ is not closed, in which case not all of the points on the future boundary of $\mathcal{J}^-(p)$ can be joined to p by a causal curve and $L^-(p)$ will be a proper subset of this boundary. I will ignore such pathologies in the discussion below.[8]

Example 2. For a future endless timelike geodesic γ in Minkowski spacetime, FEH(γ) = $\varnothing$ since γ's past light cone sweeps out the entire spacetime.

Example 3. For the future endless, hyperbolically accelerated η of Fig. 5.5 (depicting two-dimensional Minkowski spacetime), FEH(η) = H so that nothing on the opposite side of H is observable by η by means of causal signals.

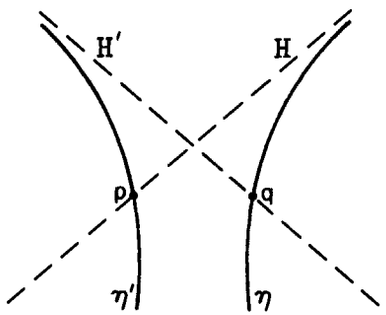

Fig. 5.5 An illustration of future event horizons in (two-dimensional) Minkowski spacetime

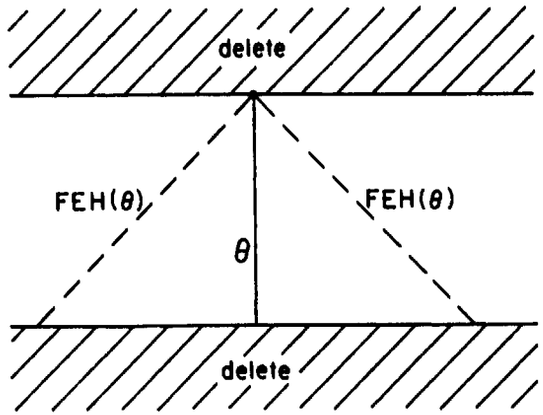

Fig. 5.6 Event horizons for geodesic observers in (two-dimensional) future- and past-truncated Minkowski spacetime

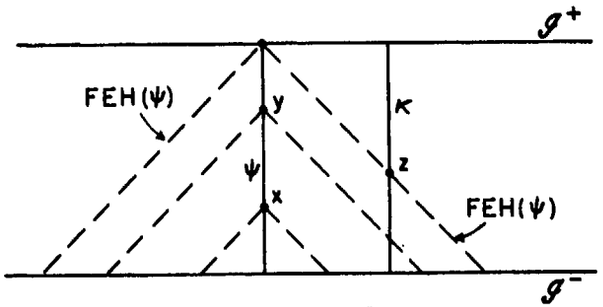

Fig. 5.7 Conformal diagram of de Sitter spacetime

Example 4. The timelike geodesic θ of future- and past-truncated Minkowski spacetime depicted in Fig. 5.6 has a non-trivial future event horizon.

Example 5. De Sitter spacetime illustrates how a geodesically complete spacetime can have future endless timelike geodesics with non-trivial event horizons (see Fig. 5.7). In this example the existence of a non-trivial future event horizon for a future endless geodesic observer is a consequence of the fact that $\mathcal{I}^+$ is spacelike. However, in asymptotically flat black hole spacetimes, non-trivial event horizons exist even though $\mathcal{I}^+$ is null (recall Fig. 3.4b).

The concept of a *particle horizon* is much murkier. Rindler applied the notion to "fundamental observers" in Friedmann–Robertson–Walker (FRW) cosmological models, that is, to timelike geodesics that represent the motions of stars, galaxies, and the like during the matter-dominated portion of the

evolution.[9] For such an observer γ the particle horizon at a cosmic instant t_0 was said to be "a surface in the instantaneous 3-space $t = t_0$, which divides all fundamental particles into two non-empty classes: those that already have been observable by $[\gamma]$ at $t = t_0$ and those that have not" (Rindler 1956, p. 663). Hawking and Ellis (1973) take the particle horizon $\mathrm{PH}(\gamma, p)$ for γ at some point $p \in \gamma$ to be "the division of particles into those seen by $[\gamma]$ at p and those not seen by $[\gamma]$ at p" (ibid., p. 128). Of course, by "seen" they mean *can* be seen via a causal signal, and by "not seen" they mean *cannot* be seen via a causal signal. Note that for any γ and γ' containing p, $\mathrm{PH}(\gamma, p) = \mathrm{PH}(\gamma', p)$. Wald (1984a) speaks of the particle horizon of γ at p as the "boundary" between world lines of particles that can be seen by γ at p and those that cannot. The definitions of Rindler and Wald seem to associate a particle horizon with a surface in spacetime, whereas the Hawking and Ellis definition involves no such surface but only a division of particles. Some concrete examples will help to illustrate some of the peculiarities and ambiguities of usage.

Example 6. In Fig. 5.4 particle β cannot be seen by γ at p, so that following the definition of Hawking and Ellis one would say that β is outside of γ's particle horizon at p. But in the standard usage, observers in Minkowski spacetime would not be said to have particle horizons since for any point p, $\mathcal{J}^-(p)$ contains the world line of any past endless timelike geodesic. This usage reflects the original concern with particles that follow timelike geodesics.

Example 7. In the conformal representation of de Sitter spacetime (Fig. 5.7), the particle following the timelike geodesic κ is beyond ψ's particle horizon at x but not at y. In this example, the existence of nontrivial particle horizons, for particles following geodesics, is a consequence of the fact that $\mathscr{I}^-$ is spacelike. But as in the case of event horizons, particle horizons can exist even when $\mathscr{I}^-$ is not spacelike.

Example 8. A more physically relevant example of particle horizons is provided by the FRW models. Because these models are the focus of the current debates about the horizon problem, they will be treated in some detail below. But first I want to make a few more general remarks about the definitions of particle horizons.

The concentration on geodesics in the typical discussion of particle horizons is a little difficult to understand. EFE entail that the energy–momentum tensor T^{ab} obeys the conservation law $\nabla_a T^{ab} = 0$. If T^{ab} represents a perfect fluid and the pressure $p = 0$—that is, we are dealing with a dust source—then it follows directly from the conservation law that the normed four-velocity V^a of the dust obeys $V^b \nabla_b V^a = 0$, i.e., the world lines of the dust particles are geodesics. But when $p \neq 0$ and, more generally, when T^{ab} is not a perfect fluid source, matter will not follow geodesics. And in any case, observers equipped with rocket engines are not confined to timelike

geodesics. On the whole it might seem preferable to drop talk about particle horizons, or else to reform usage so that a particle horizon at a point p marks the division between those particles whose world lines fall into $\mathcal{J}^-(p)$ and those whose world lines do not, regardless of whether the particle world lines are geodesics. But the standard usage, whatever its shortcomings, does have the virtue of focusing attention on a class of cases that generates the horizon problem; namely, those cases like Figs. 5.3, 5.6, and 5.7 where there are contemporaneous events that have disjoint causal pasts. When I speak of particle horizons in the context of the "horizon problem" of big bang cosmology in the sections below I will have this feature in mind.

Another potentially confusing aspect of the usual treatment of particle horizons concerns the true slogan that once within a particle horizon, always within a particle horizon; or as Rindler puts it, "particles that have at sometime been visible to $[\gamma]$ remain so forever" (Rindler 1956, p. 666). Trivially, if the world line of a particle δ falls within $\mathcal{J}^-(p)$ for some $p \in \gamma$, then for any $q \in \gamma$ such that q chronologically succeeds p, δ falls within $\mathcal{J}^-(q) \supset \mathcal{J}^-(p)$. But this truism is entirely compatible with there being a stage in δ's existence after which δ cannot causally signal to γ. This will occur when γ has an event horizon and δ has crossed it. In the Minkowski example of Fig. 5.5, η and η' are within each other's particle horizon at every moment. But after q η can no longer signal to η', and after p η' can no longer signal to η. A similar behavior can occur for geodesic observers in spacetimes with a spacelike future infinity. Thus, in a conformal diagram of de Sitter spacetime (Fig. 5.7), the geodesic κ is within the geodesic ψ's particle horizon after y; but after z, κ cannot successfully signal to ψ. Somewhat analogous behavior can occur in inflationary scenarios, as will be discussed in sections 5.11 and 5.12.

For future reference it will be helpful to give a more detailed treatment of particle horizons in the FRW models. Coordinates can be chosen so that the line element of a homogeneous and isotropic metric takes the form

$$ds^2 = a^2(t)[dr^2 + f^2(r)(d\theta^2 + \sin^2\theta\, d\phi^2)] - dt^2 \tag{5.1}$$

with $f(r) = 1/(1 - kr^2)$, where $k = +1$ (space sections $t =$ constant of constant positive curvature), $k = 0$ (flat space sections), or $k = -1$ (space sections of constant negative curvature). Because of homogeneity there is no loss of generality in focusing on a fundamental observer at $r = 0$. And because of isotropy there is no loss of generality in focusing on radial null geodesics ($d\theta = d\phi = 0$) in discussing what our chosen observer can see by optical means. Since $ds^2 = 0$ along a null geodesic, it follows from (5.1) that when at a time t our observer looks backward in time to events occurring at $t_0 < t$, his optical observations will be able to extend to a coordinate distance of

$$r(t, t_0) = \int_{t_0}^{t} \frac{dt'}{a(t')} \tag{5.2}$$

As measured at time $\hat{t}$ the proper distance from the spatial location $r(t, t_0)$ to the origin $r = 0$ is $d_r(\hat{t}) = a(\hat{t})r(t, t_0)$. Whether it is useful to choose $\hat{t}$ equal to t, or to t_0, or to something in between will depend on the application. As t_0 approaches the time of the big bang ($t = 0$), the limiting value of $r(t, t_0)$ marks the boundary of the portion of the universe that at time t is accessible to our observer by direct optical means:

$$r_H(t) =: \lim_{t_0 \to 0^+} r(t, t_0) = \lim_{t_0 \to 0^+} \int_{t_0}^{t} \frac{dt'}{a(t')} \tag{5.3}$$

Thus, a particle horizon is present for our observer at t just in case the horizon coordinate distance $r_H(t)$ is finite, which will be the case if and only if the integral in (5.3) converges.[10]

Obviously this matter cannot be decided until the temporal behavior of the scale factor $a(t)$ (aka the "radius of the universe") is specified. But to help fix intuitions, consider the mathematical example where $a(t)$ behaves as t^n. Then particle horizons will be present just in case $n < 1$.

In GTR the behavior of the scale factor is determined by EFE in conjunction with assumptions about the nature of the energy–momentum tensor. In the FRW models the symmetries of the spacetime force the energy–momentum tensor to have the form of a perfect fluid: $T^{ab} = (\mu + p)V^aV^b + pg^{ab}$, where μ is the mass density, p is the pressure, and V^a is the normalized four-velocity of the fluid. With cosmological constant $\Lambda \leq 0$ and with $\mu + 3p > 0$ (as will certainly be the case, for example, with positive mass density and non-negative pressures), EFE entail that there is an initial singularity. And for $p \geq 0$ the behavior of $a(t)$ is such that the integral in (5.3) converges (see section 5.11).

The FRW metric is conformally flat. In the case of a flat spatial geometry ($k = 0$) this means that new coordinates can be chosen such that the line element takes the form

$$ds^2 = \Omega^2(\tilde{t})[dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2) - d\tilde{t}^2] \tag{5.4}$$

The case of particle horizons then corresponds to the divergence of the conformal factor $\Omega(\tilde{t})$ as $\tilde{t} \to 0^+$.[11] Since causal properties are not affected by a conformal transformation, the causal features of the FRW models with particle horizons can be represented as in Fig. 5.8, which gives a much better intuitive feeling for particle horizons than do the convergence properties of the integral in (5.3). The price to be paid is that Fig. 5.8 badly distorts spatiotemporal distances.

5.5 What is the horizon problem?

Suppose that our universe is an FRW big bang universe with particle horizons. We cannot see all the way back to the big bang with either an optical or a radio telescope because the regime prior to the time t_d ($\approx 10^{13}$ s)

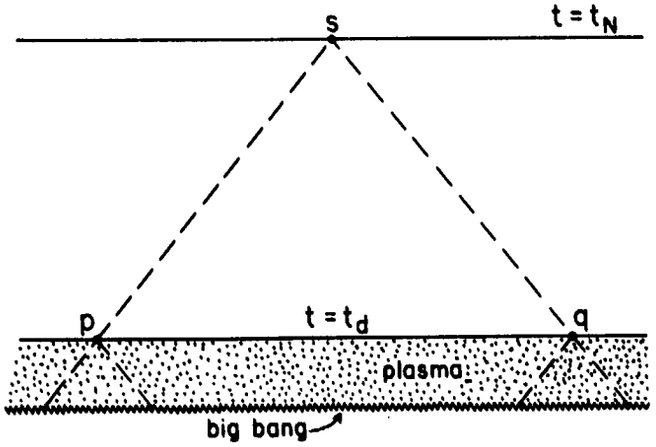

Fig. 5.8 The standard big bang model in conformal representation

of the decoupling of matter and radiation is opaque to such means of observation. When at $t = t_N$ ("now" $\approx 10^{18}$ s) we look backwards in time towards the decoupling, we observe microwave radiation, a remnant of the big bang, that is remarkably homogeneous and isotropic. (Recently, small-scale fluctuations in the cosmic microwave background radiation (CMBR) have been detected, but this does not affect what follows.) But the FRW model implies that for directions in space with a sufficiently large angular separation, the events at the time t_d of decoupling (e.g., p and q in Fig. 5.8) have no common causal past.

The conjunction of this causal disjointness and the isotropy of the CMBR is variously said to be "quite puzzling" (Hakim 1984), a "paradox" (Barrow and Silk 1980), a "major mystery" (Rees 1972), and "philosophically unsatisfactory" (MacCallum 1979). Several different research programs heading off in different directions have been proposed to deal with this horizon problem. Some of them will receive attention in due course. But before turning to attempts to resolve the problem, it is important to get a fix on what the problem is supposed to be.

Philosophers of science who are steeped in the literature on scientific explanation and causality will be tempted to give a quick answer. In everyday life and science, it is usual to search for a common cause explanation for a correlation between distant events. But in the case at issue no such explanation is possible since there is no common causal past for the events in question. To explore this answer we need to know a bit more about Reichenbach's principle of common cause.

5.6 Reichenbach's principle of common cause

The principle of common cause (PCC) was enunciated in Reichenbach's posthumously published book, *The Direction of Time* (1971). In the succeeding

years the PCC gained an important niche in the philosophy of science: it turns up in treatments of scientific explanation, causality, and realism, and it also enters discussions of the foundations of biology and quantum mechanics (see Salmon 1984; van Fraassen 1980; Sober 1988; Arntzenius 1992, and the references given therein). And yet both the status and the content of the PCC are the subject of vigorous debate.

Reichenbach's intentions in positing the PCC are not pellucid, but it seems that he wanted to frame a principle that would give both relativity theory and quantum mechanics their due. The former he took to imply no action at a distance and the latter he took to dethrone determinism. In keeping with the latter it is hopeless to look for deterministic causes for events, while in keeping with the former (Reichenbach thought) we ought to find a common cause, albeit probabilistic in nature, that explains the correlation between distant events. More specifically, if $A(p)$ and $B(q)$ are events that occur at relatively spacelike locations p and q and $\Pr(A(p) \,\&\, B(q)) \neq \Pr(A(p)) \times \Pr(B(q))$, then a *probabilistic common cause* is an event $C(w)$ such that $w \in \mathcal{J}^-(p) \cap \mathcal{J}^-(q)$ and $\Pr(A(p) \,\&\, B(q)/C(w)) = \Pr(A(p)/C(w)) \times \Pr(B(q)/C(w))$, or equivalently (assuming no divisions by 0) $C(w)$ *screens off* $A(p)$ from $B(q)$ and vice versa in that $\Pr(A(p)/C(w) \,\&\, B(q)) = \Pr(A(p)/C(w))$ and $\Pr(B(q)/C(w) \,\&\, A(p)) = \Pr(B(q)/C(w))$.

Quantum mechanics, which was one of the two prompters of the PCC, implies two ironies. First, there are quantum states, such as the singlet state for two spin 1/2 particles, that involve a strict correlation between events (such as measurements of the spins of the particles along chosen axes) which can be arranged to be relatively spacelike. For such strict correlations a screener-off is possible only if the conditional probabilities on the common cause are 0 or 1, which is a return to determinism or something very near it (see van Fraassen 1980). Second, even when the correlations are not strict, the PCC clashes with quantum mechanics. For the existence of a probabilistic common cause implies a set of probabilistic inequalities—for instance, the Bell–Clauser–Horne inequalities—that are provably violated by quantum statistics (see Clauser and Horne 1974).

One reaction to these facts would be to interpret the PCC as asserting that either correlated relatively spacelike events have a common cause or else they are connected by a direct causal link (see Arntzenius 1992). This reading turns the PCC into a purely classificatory principle whose only function is to distinguish between two ways relatively spacelike events can be causally linked—either indirectly via a common cause or directly through some non-local action. Parts of Reichenbach's *The Direction of Time* and his earlier book *Philosophy of Space and Time* (1958) suggest that he had a stronger interpretation in mind, one that gives the PCC a normative dimension, for Reichenbach apparently thought that relativity theory prohibits non-local action. Relativistic quantum field theory (QFT) provides a counterexample to the normative reading that says that relativity theory is incompatible with

any non-local action *as defined by a failure of screening off*. QFT is the proper arena for discussing the PCC in connection with quantum phenomena; for the spacetime background of this theory is Minkowski spacetime (as opposed to Newtonian spacetime for ordinary quantum mechanics), and in addition QFT (unlike ordinary quantum mechanics) allows operators to be associated with spacetime regions. Thus, in QFT one can make precise sense of the notion, which one can only indicate by hand-waving gestures in ordinary non-relativistic quantum mechanics, of measurements made at relatively spacelike locations. Violations of the Bell inequalities and, hence, of the PCC can be arranged in this setting (see Landau 1987; Summers and Werner 1977a, b). But this failure of the PCC does not signal any sort of non-locality that is in conflict with relativity theory, as is made manifest by the relativistic invariance of QFT.[12] Of course, one could still insist on the weaker, classificatory reading of the PCC and conclude that quantum phenomena do involve some sort of non-local features. And in seeming concert with this reading there is a large literature that consists of hand-wringing and moaning about quantum non-locality. There certainly are non-local features to QFT, and some of them are surely puzzling. For example, even with the axiom that operators associated with relatively spacelike regions commute (or anticommute) one still has the Reeh–Schleider theorem which asserts that operating on the vacuum state with polynomials of operators associated with some bounded region of spacetime generates a dense set of states in the Hilbert space. One would like to understand better such non-local features. But in carrying out this task the philosophy underlying the PCC is useless—indeed, less than useless, for it suggests that these non-local features indicate a lurking conflict with relativity theory, whereas relativistic invariance has been built into the construction of QFT ab initio.

Even if we leave aside the mysteries of the quantum domain, the status of the PCC is still open to challenge. A forceful illustration has been provided by Arntzenius (1992), who shows that for generic, time homogeneous, Markov processes a screener-off will generally not exist. In addition, Forster (1986) and Arntzenius (1992) argue persuasively that there are a variety of equilibrium correlations which do not call for a common cause explanation.[13] Of course, even if the PCC cannot be defended as a general principle, it may nonetheless help to pinpoint what is problematic about the horizon problem in general relativistic cosmologies with particle horizons. It is to that matter I now turn.

5.7 Particle horizons and common causes

It is not easy to bring Reichenbach's PCC to bear on relativistic cosmological models. As a first stab, it might seem that the PCC is necessarily satisfied in this setting since GTR is a deterministic theory, which means that the relevant probabilities are always 0 or 1. This is too quick and too crude. It is too quick

because it is not evident that the probabilities can be made 0 or 1 by conditioning on events in the common causal past (more on this anon). It is also too crude because it does not address the nature of probabilities in GTR. At the metalevel probabilities can enter in terms of measure theoretic arguments concerning classes of solutions of EFE (see section 5.9); but this is irrelevant to the PCC. At the object level, probabilities can enter in one of two ways. First, they can be epistemic probabilities, representing our assessments of how likely the available evidence makes the occurrence of the events in question. Again, however, the PCC was not intended to apply to such measures of our ignorance. Second, the probabilities may enter by introducing some relativistic version of classical statistical mechanics, e.g., the relativistic Boltzmann equation. But the central and unsolved mystery of non-quantum statistical mechanics is what these probabilities mean and how they are to be justified. The upshot is that any attempt to apply Reichenbach's PCC to classical GTR is muddied by one of the most contentious foundation problems in physics.

Can we not cut through some of these difficulties and say at least the following? There is a perfectly good intuitive sense—even if we cannot make it precise by the use of unproblematic probability assertions—in which relatively spacelike events in general relativistic cosmological models can be correlated. In models with particle horizons these events may have no common causal past. Hence, there is a conflict with the spirit if not the letter of the PCC, and this conflict helps to illuminate the widely expressed queasiness about particle horizons. This answer is overly optimistic. The conflict provides illumination only if one accepts the spirit of Reichenbach's assumptions about common causal explanations; but the most straightforward attempt to translate these assumptions into the setting of GTR leads to an incoherent view of explanation. The reader who digested the discussion of deterministic prediction in section 5.3 will have anticipated the point, but it merits explicit elaboration here.

In the context of general relativistic theories, to construct a dynamical explanation of the state in some region $X \subset M$ spacetime M, g_{ab} requires the choice of a spacelike hypersurface Σ such that $X \subset D^+(\Sigma)$. This is so whether X is a connected region R or the disjoint union of two relatively spacelike regions R and R'. The specification of the relevant initial data on Σ together with the coupled Einstein–matter equations entail the state on X. In the philosophical jargon, we have a paradigm case of Hempelian deductive–nomological explanation. In some instances the existence of naked singularities or other pathologies in the spacetime structure may prevent the choice of such an Σ (see chapter 3). But in general the existence of particle horizons is *not* such a pathology; and in particular, in the FRW models the presence of particle horizons does *not* entail the non-existence of such a Σ since the FRW models can be partitioned by Cauchy surfaces. Conversely, even when there are no particle horizons, it will generally not suffice for a dynamical explanation of the states on relatively spacelike regions R and R'

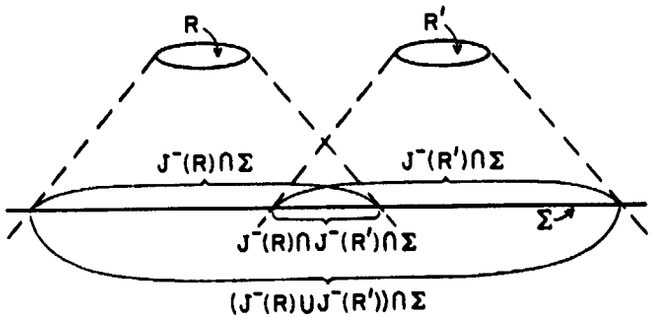

Fig. 5.9 The common causal past of two relatively spacelike regions

to choose a Σ that belongs to the common causal past $\mathcal{J}^-(R) \cap \mathcal{J}^-(R')$, for as already noted above it is typically the case that $D^+(\mathcal{J}^-(R) \cap \mathcal{J}^-(R')) = \mathcal{J}^-(R) \cap \mathcal{J}^-(R')$. It was also noted that there are exceptions (e.g., where $\mathcal{J}^-(R) \cap \mathcal{J}^-(R')$ contains a Cauchy surface); but these exceptions are truly exceptional among general relativistic cosmological models.

In sum, the problem of the "horizon problem" cannot be that a paradigm deductive–nomological explanation cannot be constructed in the presence of particle horizons, for that complaint is patently false in typical cases. Nor can the complaint be that in the presence of particle horizons such an explanation fails to be a good explanation because it fails to be a common cause explanation; for in typical cases—even *without* particle horizons—it is impossible to give a dynamical explanation of the type under discussion purely by reference to the common causal past.

By way of defense of common cause explanations let us suppose that a spacelike hypersurface Σ is chosen large enough with respect to the relatively spacelike regions R and R' and that the spacetime lying between Σ on the one hand and R and R' on the other is non-pathological; that is just to say that $R \subset D^+(\mathcal{J}^-(R) \cap \Sigma)$ and $R' \subset D^+(\mathcal{J}^-(R') \cap \Sigma)$ (see Fig. 5.9). Even though $(R \cup R') \not\subset D^+(\mathcal{J}^-(R) \cap \mathcal{J}^-(R') \cap \Sigma)$, it is reasonable to think that in typical cases where $\mathcal{J}^-(R) \cap \mathcal{J}^-(R') \cap \Sigma$ is, in some appropriate sense, a large fraction of $(\mathcal{J}^-(R) \cup \mathcal{J}^-(R')) \cap \Sigma$, the heart of the dynamical explanation of the joint state on R and R' lies with the initial data on the common causal past portion $\mathcal{J}^-(R) \cap \mathcal{J}^-(R') \cap \Sigma$ of Σ. In these circumstances there is a deterministic deductive nomological explanation approximating a common cause explanation. Particle horizons pose a problem because they tend to block the satisfaction of the large fraction condition.

All of this is true enough but not really to the point. Suppose that because of the presence of particle horizons $\mathcal{J}^-(R) \cap \mathcal{J}^-(R') \cap \Sigma = \varnothing$ for any connected partial Cauchy surface Σ but that there exists a connected partial Cauchy surface Σ such that $(R \cup R') \subset D^+((\mathcal{J}^-(R) \cup \mathcal{J}^-(R')) \cap \Sigma)$. So we have an explanation, in terms of the dynamical evolution of the appropriate initial data on $(\mathcal{J}^-(R) \cup \mathcal{J}^-(R')) \cap \Sigma$, of the joint state on R and R', an

explanation which is not in any sense a common cause explanation. Our problem is this: *What precisely is it about such a dynamical explanation that makes it deficient or unsatisfying as an explanation?* Discussions surrounding the PCC hint that something must be wrong with the explanation since by hypothesis it cannot rely on common causes. But the hints are far from articulate.

Knowing that we cannot rely on the PCC to solve our problem, there appears to be no other recourse than to investigate in detail the physics associated with particle horizons in order to try to pinpoint what it is that makes models with a particle horizon problematic. What this investigation uncovers can perhaps be identified with the valid core of the PCC as applied to particle horizons.

5.8 Diagnosing the bellyache: Electromagnetism

In trying to diagnose the bellyache involved with particle horizons it will be helpful to reconsider in more detail the remarks about electromagnetism made in section 5.2.

In four-dimensional Minkowski spacetime Maxwell's equations entail that the components of the vector and scalar potentials for the electromagnetic field all obey the inhomogeneous wave equation

$$\Box\Phi(x, y, z, t) = f(x, y, z, t) \tag{5.5}$$

where $\Box =: \eta^{ab}\nabla_a\nabla_b$, ∇_a is the derivative operator associated with the Minkowski metric η_{ab}, and f describes the source distribution.[14] Any solution of (5.5) can be written in the Kirchoff retarded representation, which has the form

$$\Phi(Q, t) = \int_V \text{ret} + \int_A \text{ret} \tag{5.6}$$

The volume integral in (5.6) is $\int_V \frac{[f]}{r}\, dV$ where V is a volume r containing Q, r is the distance from Q to the volume element, and $[f]$ means that f is to be evaluated at the retarded time $t' = t - r$. In plain English, the volume integral represents the contribution to the field $\Phi(Q, t)$ coming from sources at spacetime locations where the past light cone $L^-(Q, t)$ cuts the world lines of the sources. The second term on the right-hand side of (5.6) is a surface integral. It gives a contribution only from points that lie at the intersection of $L^-(Q, t)$ with the past time slice corresponding to $t' = t - r$. (This is an expression of *Huygens' principle*, which means that the effects of Φ propagate cleanly at exactly the speed of light. This principle fails if the dimension of the space is, say, two.) This surface integral gives contributions from (i)

sources outside of V, and (ii) source-free radiation. If there are only a finite number of sources (here, charged particles) then as V is expanded eventually all sources will be picked up and (i) will cease to contribute. If we further postulate that there is no source-free radiation coming from infinity (*Sommerfeld radiation condition*), then in the limit in which V expands without bound, or correspondingly the retarded time t' is pushed back towards $-\infty$, (ii) ceases to contribute. We are left with the Linard–Wiechart potential

$$\Phi(Q, t) = \int_V \frac{[f]}{r}\, dV \tag{5.7}$$

familiar from texts on electromagnetism.

Now suppose that particle horizons are introduced by truncating Minkowski spacetime in the past, as in Fig. 5.3. Then (5.7) cannot hold. This is just a way of reexpressing the point about constraint equations made in section 5.2. Because of the presence of the charged particle δ, the value of Φ at $p = (Q, t)$ cannot be zero even though $L^-(Q, t)$ never cuts δ. This means in turn that the Sommerfeld radiation condition cannot be satisfied; no matter how far the retarded time is pushed towards the beginning of time at 2000 B.C., there will still be source-free radiation entering the volume. (This is not to say that the original solution to Maxwell's equations is no longer a solution; indeed, any solution on full Minkowski spacetime remains a solution when restricted to past-truncated Minkowski spacetime, as follows from the local nature of Maxwell's equations. The point is that the restriction of a solution, although still a solution, may not admit a representation with the desired properties, such as no incoming radiation.)

Needless to say, past-truncated Minkowski spacetime is a highly artificial illustration of particle horizons. So it is natural to wonder about the fate of the Sommerfeld radiation condition in non-artificial cosmological models with particle horizons. The matter is complicated, in part because in general Huygens' principle does not hold for curved spacetimes and in part because in this general setting it is not easy to give precise mathematical expression to the idea that there is no incoming source-free radiation. However, the investigations of Penrose (1964) and Ellis and Sciama (1972) indicate that the Sommerfeld condition cannot be satisfied in generic cosmological models with particle horizons arising from a spacelike $\mathscr{I}^-$.[15]

The failure of the Sommerfeld radiation condition will be a source of consternation to those who share the Machian intuition that all physical effects must be explained in terms of sources in the form of ponderable bodies. Einstein (1916) himself was of this persuasion when he wrote the paper that put GTR in its final form.[16] Perhaps the wide appeal of this intuition helps to explain why cosmological models with particle horizons are thought to be objectionable. But the validity of the objection is another matter. The failure of theories of modern physics to conform to Machian intuitions may indicate

that there is something wrong with these theories. But one should also be prepared to consider the alternative that these intuitions need retraining. Certainly Einstein's special and general theories of relativity promote the view that fields are entities in their own right and are ontologically as basic as particles. And in quantum field theory the now dominant point of view is that all particle-like behavior is to be explained in purely field theoretic terms and that there are many circumstances where the particle concept is not useful or even applicable.

It should also be emphasized that implementing the condition of no source-free radiation comes at a price; namely, it assumes an asymmetry of time. To appreciate the point, return to the simpler context of STR and Minkowski spacetime. The retarded representation (5.6) is just that—it is not a particular solution of (5.5) but rather a representation of a general solution. But there are many other representations as well. The advanced representation, for example, is obtained by evaluating the integrals in (5.6) at the advanced time $t + r$. And in addition there are any number of linear combinations of advanced and retarded representations. Sticking to the pure retarded and advanced representations, we have

$$\int_V \text{ret} + \int_A \text{ret} = \int_V \text{adv} + \int_A \text{adv} \tag{5.8}$$

So if we require that $\int_A \text{ret} = 0$, it follows that

$$\int_A \text{adv} = \int_V \text{ret} - \int_V \text{adv} \tag{5.9}$$

And, as Sciama (1963) notes, $\int_A \text{adv}$ will not in general vanish. Perhaps this temporal asymmetry will be welcomed, and perhaps it can be justified, e.g., by showing that under certain cosmological conditions the advanced representation cannot be valid (see Sciama 1963; Hawking 1965), thereby providing at least part of the solution to the problem of the direction of time. This is not the place to tackle this knot of contentious issues. The point I want to emphasize is simply that the seemingly innocuous condition of no source-free radiation carries with it some heavy baggage.

To summarize, if we subscribed to the prejudice that in electromagnetism and gravitation all effects must be due to causal propagation from sources in the form of ponderable bodies, then we would have at least a partial account of what is objectionable about cosmological models with particle horizons. But this prejudice is nothing more than a prejudice.

5.9 Diagnosing the bellyache: Cosmic background radiation

How are we to understand pronouncements such as "the existence of [particle] horizons is a fundamental obstacle to any dynamical explanation"

of the homogeneity of the CMBR (Hartle 1983, p. 79) or that because of the presence of particle horizons in the standard big bang models it is "essentially impossible" to account for the uniformity of the CMBR as having evolved "due to a physical process operating in the early Universe" (Turner 1987, p. 226)? In *The Early Universe* Börner at first seems to want to dismiss talk about a horizon problem.

> It does not seem fair to speak of "problems" in the context of the standard big bang model. In any solution of a differential equation there are certain specific properties of the initial data. If we compute backwards in time we just find the initial data that are responsible for the state of affairs as we see it now. (Börner 1988, pp. 274–275)

But he goes on to add

> If we look at the present state of the universe as a consequence of certain initial data we might feel a bit uneasy, if the initial data have to be extremely specific. . . . As physicists we would feel more at ease if we could find an understanding of such specific conditions in terms of physical processes.[17] (ibid., p. 275)

Similarly, Turner (1987) notes that the uniformity of the CMBR "can be accommodated by the standard model, but seemingly at the expense of highly special initial data" (p. 227).

There are two contentions here. The first is that a physical theory that postulates "special initial conditions" is somehow lacking or inadequate. The second is that standard models of the big bang are forced to posit special initial conditions. The first contention calls for careful evaluation since it gets to the heart of issues about what makes a good scientific explanation. But before turning to the evaluative task, I want to examine the second contention.

A crude argument for the second contention goes as follows. If the universe starts in an inhomogeneous and anisotropic state in the presence of particle horizons, then it cannot achieve uniformity in a reasonable time; for distant parts of the universe cannot causally interact and, hence, there is no physical mechanism for producing uniformity within the required time span. From the previous sections we know that this argument is too quick. For in inhomogeneous and anisotropic universes, particles beyond each other's particle horizons do interact in the sense that each feels electrical and gravitational forces due to the unseen presence of the other, and these forces might conceivably provide a mechanism for helping to achieve uniformity. What is true in classical relativistic physics is also true in relativistic quantum physics. QFT entails the existence of correlations between relatively spacelike regions, even if these regions have no common causal past. Let O_1 and O_2 be open regions of spacetime and let $\mathscr{A}_1$ and $\mathscr{A}_2$ denote the local algebras of observables (self-adjoint operators) associated with O_1 and O_2 respectively. $A_1 \in \mathscr{A}_1$ and $A_2 \in \mathscr{A}_2$ are said to be correlated with one another in the quantum state Ψ just in case $\langle\Psi|A_1A_2|\Psi\rangle \neq \langle\Psi|A_1|\Psi\rangle\langle\Psi|A_2|\Psi\rangle$. When O_1

and O_2 are relatively spacelike, a standard causality axiom of relativistic QFT asserts that A_1 and A_2 commute. Nevertheless, A_1 and A_2 may be correlated; indeed, the Reeh–Schleider theorem[18] entails that for any O_1 and O_2, no matter how far apart, there are $A_1 \in \mathscr{A}_1$ and $A_2 \in \mathscr{A}_2$ such that A_1 and A_2 are correlated in the vacuum state (see Wald 1992).[19] It is conceivable that the correlations between regions that lie beyond each other's particle horizons might be strong enough to have a discernible effect in these regions (see Wald 1992). A perhaps more effective mechanism for reducing anisotropy in the CMBR is at work in compact $k = -1$ FRW models where the sensitive dependence on initial conditions of the spacetime geodesics leads to a flow on the surfaces of homogeneity that is strongly chaotic (see Ellis and Tavakol 1994).

It seems to me implausible that such trans-particle horizon mechanisms could play a major role in smoothing out inhomogeneities and anisotropies. But the point is that some non-negligible role cannot be ruled out a priori as some of the thinking behind the PCC would suggest. In any case a more sophisticated argument is needed to establish that in the presence of particle horizons no mechanism can succeed in smoothing out inhomogeneities and anisotropies in the initial state. The argument might take the form of detailed calculations for various potential smoothing mechanisms. Or in the case of classical GTR it might take the form of a measure theoretic examination of solutions of EFE. According to Turner (1987), an argument of the latter type has been provided by Collins and Hawking.

In 1973 Collins and Hawking "pointed out that the set of initial data which evolve to a Universe such as ours is of measure zero providing that the stress energy in the Universe has always satisfied the strong and dominant energy conditions" (Turner 1987, p. 227). More specifically, what Collins and Hawking (1973) showed is this: assuming that EFE are satisfied with $\Lambda = 0$ and that matter obeys the dominant energy condition and the positive pressure criterion,[20] then there is no open set of initial data which leads to cosmological models that eventually approach isotropy and which intersects the subspace of homogeneous initial data. As Collins and Hawking note, this conclusion does not by itself rule out the possibility that there is an open set of *in*homogeneous initial data that leads to homogeneity and isotropy at late times, but it makes the possibility seem unlikely since presumably inhomogeneities in the initial data are more likely to lead to anisotropy than to isotropy.

What the powerful result by Collins and Hawking raises in the first instance is not a horizon problem but a uniformity problem. Since initial data and cosmological models satisfying EFE are in one–one correspondence, the Collins–Hawking theorem shows that (in all plausibility) cosmological models that achieve uniformity even at late times are very rare in the space of all models satisfying EFE and the energy and pressure conditions. To lay the blame for the uniformity problem at the doorstep of particle horizons requires further results along the following lines. Since the big bang is the focus of concern, one would like to single out a subspace of models that can be said

to begin with a big bang. Next one would like to show that in the appropriate subspace topology there is an open set of such big bang models which satisfy EFE and energy conditions and which eventually achieve homogeneity and isotropy. And finally one would try to show that within this open set there is no open subset of models with particle horizons. In the light of such results one could reasonably say that the uniformity of the universe is not an a priori problem in big bang models but that it becomes a problem if particle horizons are present. This is one more example of how tacit assumptions are needed to make the horizon problem into a genuine problem. Nor is it obvious that the tacit assumptions are warranted; for instance, we will see in section 5.12 that there is a natural way to assign a measure to initial conditions such that uniformity is an a priori problem and not a problem of particle horizons.

For the sake of discussion I want to suppose now that we have in hand results that indicate that particle horizons make uniformity at late times a rare or improbable feature. It still remains to give a more detailed diagnosis of this complaint against particle horizons. I will offer a twofold classification for making the complaint more specific. The classification is based on two different ways of faulting a putative explanation on the basis of improbability considerations. The first way of faultfinding is the least searching, for it agrees at the outset that *if* the hypotheses of the putative explanation were true, then a satisfactory explanation would have been provided. But it goes on to claim that in all probability the putative explanation involves a false assumption.

A neutral example may help to make the nature of the first complaint more concrete. Suppose that the serious candidates for explaining the origin of the earth's moon are: H_1: condensation/accretion (i.e., the moon was formed at the same time and by the same process by which the planets condensed out of primordial matter); H_2: fission (i.e., the moon was split off from the earth by an impact or some other mechanism); and H_3: capture (i.e., the moon came from outside our solar system and was drawn into its present orbit by the earth's gravitational field). It will be agreed, I think, that *if* H_3 were true, then spelling out the details would provide a satisfactory explanation of the origin of the moon. The difficulty is that the set of initial conditions that would lead to capture is of measure zero. Adopting Bayesian terminology, it would seem plausible to set the prior probability of H_3 very low.[21] But notice that one can easily imagine additional evidence that would boost the posterior probability of H_3 to a respectable value and thus remove the objection at issue. For example, the first astronauts to land on the moon might have brought back rock samples that strongly indicated an extrasolar origin for the moon. In this case the probability of H_1 and H_2 would be dramatically lowered and the probability of H_3 would be correspondingly increased. Similarly, on the basis of the Collins–Hawking result (and hypothesized supplementary results) one might be inclined to assign a low prior probability to the standard big bang account of the uniformity of the CMBR. But again one can easily imagine evidence that would substantially shift the posterior probability in favor of the standard model. For instance,

neutrino and gravitational wave detection might improve to the extent that we are able to probe conditions at times well before the decoupling time t_d, and the probes might reveal that at these earlier times the conditions were homogenous and isotropic. Then once again the first form of the improbability objection would fall away.[22]

The second way of faultfinding claims to locate a deeper flaw in the explanation offered by the standard big bang model. The objection is not that the Collins–Hawking result and hypothesized related results show that the standard explanation has a low prior probability but rather that the explanation only works for a negligible range of initial data. So even if neutrino and gravitational wave astronomy were to indicate that the early universe was in fact uniform, the objection would still stand that the explanation lacks robustness. This interpretation is suggested by the above quotation from Börner. It is much more explicit in Misner's (1968, 1969) program of "chaotic cosmology" which had the goal of showing how neutrino viscosity and other mechanisms would damp out inhomogeneities and anisotropies of an arbitrary initial state.

Robustness of explanation has a nice ring to it. But the general claim that the lack of robustness of a putative explanation (in the sense that it works only for very circumscribed initial data) automatically discredits or diminishes the value of the explanation seems to me to be unsustainable. The above example of the origin of the earth's moon and other similar examples, where the true explanation may rest on very special initial conditions, suffice to make the point. It remains open that there is something specific to the lack of robustness in the standard big bang explanation of the uniformity of the CMBR that undermines its value as an explanation. But I have found nothing in either the physics or the philosophy of science literature that provides anything like an argument for—rather than a mere assertion of—this more cautious claim. Nor is it clear to what extent even this more cautious claim is actually shared by the astrophysics community. It is certainly true, as will be described in the next section, that the horizon problem has given rise to an impressive variety of research programs. But this might not be an indication that the advocates of these programs subscribe to the claim in question but only that, in line with the first and weaker form of the improbability complaint, they are laying their bets that the standard big bang model is false but would be willing to withdraw their bets and their objections to the standard model if it were revealed that the very early universe was uniform. Here I will have to leave it to competent sociologists of science to conduct the relevant interviews to decide this matter.

In closing, I note that a valid or at least defensible core to the PCC has been located for the cosmological setting. When particle horizons are present and a common cause explanation of the correlation between relatively spacelike events is impossible because these events have no common causal past, then the postulation of special initial conditions is required to achieve an explanation.

5.10 Strategies for solving the horizon problem

What follows is not supposed to be an exhaustive classification of attempts to solve or dissolve the horizon problem but only serves the purpose of alerting the reader to the kinds of reactions that have been explored in the literature. The collection is motley, and some of the proposed solutions may strike one as far-fetched. But these features serve to reinforce how seriously the horizon problem is taken—for good reason or no—and to what lengths physicists are prepared to go in order to resolve it. I begin with three strategies that I take to lie outside the mainstream of current opinion as expressed in the astrophysics literature.

Anthropic solution

The title of Collins and Hawking's 1973 paper is "Why Is the Universe Isotropic?" In the last sentence of the article they state their answer: "Because we are here." More specifically the idea is that only in universes that approach isotropy can galaxies be expected to form. Since galaxies are a necessary condition for life as we know it, we should not be surprised to find that we live in a universe that is (on the appropriate scale) isotropic—if it weren't, we wouldn't be here to ask the question. Whether such anthropic explanations are genuine explanations or only soothing nostrums is part of a lively debate that is not appropriate to enter here.

The Penrose conjecture

Penrose (1979, 1986, 1988, 1989a) has argued that there is a fundamental misunderstanding involved in the hand-wringing about how special the initial conditions in the standard big bang model would have to be in order to accommodate the presently observed uniformity of the CMBR. In his opinion the operation of the second law of thermodynamics requires that the initial entropy of the universe be small, and this in turn requires that the big bang be highly constrained. Penrose's working hypothesis for expressing the constraint is the vanishing of the Weyl conformal curvature (or more precisely, the smallness of the Weyl curvature in comparison with the Ricci curvature).[23] His conjecture is that such constraints must be grounded in some yet to be discovered time asymmetric laws that lie at the juncture of quantum theory and GTR.

A universe with handles

Hochberg and Kephart (1994) postulate that at the Planck time t_P ($\approx 10^{-44}$ s) the universe was riddled with wormholes. To make the wormholes traversable and, thus, to permit causal connections among otherwise causally unconnected

regions, the weak energy condition must be violated at the wormhole mouths.[24]

I turn now to more mainstream attempts to solve the horizon problem. The first has already been mentioned in section 5.8.

Misner's chaotic cosmology

Type IX Bianchi models are standard general relativistic cosmological models which are homogeneous but non-isotropic. In some of these models one spatial direction is horizon free. The changing of this direction often enough as time goes on (the so-called 'mixmaster universe') may explain the decay of anisotropy (Misner 1969). One can then go on to study mechanisms that will dissipate inhomogeneities, hopefully arriving at an explanation of how a generic initial state can smooth itself out enough to fit current observations. Two criticisms have effectively undercut interest in this program: the probability of mixmaster behavior is low in Type IX models, and the amount of dissipation of anisotropy is not sufficient to account for present observations (see Stewart 1968; Collins and Stewart 1971; MacCallum 1971, 1979).

Getting rid of particle horizons

The most effective way to resolve the horizon problem is to get rid of the feature that gives rise to the problem. The tactics which have been proposed for implementing this strategy can be grouped into two categories. (a) *New physics before t_p.* (i) Zee (1980) proposed a symmetry-breaking mechanism involving a scalar field whose presence effectively weakens the gravitational "constant" as one goes backward in time. In terms of the discussion of section 5.3, the behavior of the scale factor in the Zee model near the big bang is $a(t) \sim t$ so that the integral (5.3) diverges. This proposal has been criticized by Linde (1980) and Sato (1980), and subsequently defended by Pollock (1981). (ii) Anderson (1983, 1984) and Hartle (1983) have explored quantum gravity effects as a means of eliminating particle horizons. Here the mechanism is the back-reaction of particle creation caused by the presence of inhomogeneities and anisotropies in the early universe. (iii) Akdeniz et al. (1991) studied a string-dominated early universe model that evolves into a radiation-dominated universe. As in (i), $a(t) \sim t$ at early times. (b) *No new physics but topological identifications.* Ellis (1971) and Ellis and Schreiber (1986) have explored the possibility of partially eliminating particle horizons by identifying points in such a way that after some chosen time observers can see all the way round the universe. Such "small universes" are easily constructed from the FRW models. Thus, for example, in the $k = 0$ case the points on the spatially flat space sections can be identified modulo some triple of distances d_1, d_2, d_3 to give a torodial spacetime topology in which one returns to the same spatial location by moving in a straight line a distance d_1 (respectively, d_2, d_3) in the x (respectively, y, z) direction. With two spatial dimensions

suppressed, the causal properties of such a model are similar to those of a version of Fig. 5.1 truncated in the past so as to simulate the big bang. In contrast to (a) the elimination of particle horizons is partial rather than total; that is, for any given t, the identification scale can be chosen small enough so as to assure that at t an observer can see around the universe (many times, if you like), but it is not true that the identification scale can be chosen so that at no time $t > 0$ is there a particle horizon. The partial character of the elimination is of no consequence if one is concerned only with explaining observations made now. But as discussed in the next section, improvements in technology may allow us to probe further and further back in time when particle horizons are present. And if one is concerned with the strong form of the horizon problem as distinguished in section 5.9, then it has to be shown how the present observation of a uniform CMBR is compatible with a generic non-homogeneous and non-isotropic initial state. Making small universes out of generic universes by the sort of identification procedure used in the $k = 0$ FRW models is in general not possible—invariance under a discrete group of isometries is required. But lumpy or inhomogeneous small universes certainly can be allowed, and one can study how uniformization of the CMBR is achieved in such multiply connected models (see Ellis and Schreiber 1986).

Making the particle horizons effectively small

Even if particle horizons exist at the time t_d of decoupling of matter and energy, something approaching a deterministic dynamical explanation in terms of a common cause may be possible for two relatively spacelike events that lie on $t = t_d$ and that are now visible to us if the common causal past $J^-(p) \cap J^-(q)$ of the spacetime locations p and q of these events is in some appropriate sense a large fraction of $J^-(p) \cup J^-(q)$ for a time near the big bang. At least two schemes have been devised to accomplish this aim. (a) The first involves the use of the Brans–Dicke theory (BDT). BDT is a generalization of classical GTR to include an adjustable parameter ω and a scalar field φ that serves as an additional source for the gravitational field. With φ constant and ω set large enough, BDT replicates the predictions of GTR. But if in the early universe one tunes the available parameters of a Brans-Dicke model, the desired condition on the common causal past can be achieved (see Dominici et al. 1983). (b) The second and more popular approach involves inflationary cosmology. Here the standard big bang model is supposed to hold up until the Planck time t_p. But at some later time t_i a new physical process not taken into account by the standard model is supposed to enormously increase the expansion of the universe so that a region initially the size of an atom at the beginning of inflation grows through inflation to a size bigger than the presently observable portion of the universe. This inflation is claimed by its proponents to solve the horizon problem. That claim will be examined in due course. But first the inflationary model needs to be examined in more detail.

5.11 Horizons in standard and inflationary models

Inflationary scenarios appeal to theories of elementary particles that unify the strong, weak, and electromagnetic forces (so-called grand unified theories or GUTs).[25] It is postulated that during the initial hot era immediately following the big bang, the symmetry which unites these forces is unbroken and consequently the predictions of the inflationary model for the expansion of the universe agree with the standard big bang model. At a later stage, however, the predictions of the two models radically diverge. To discuss these predictions, EFE (without cosmological constant) are applied to the FRW line element (5.1)[26] to obtain two ordinary differential equations for the scale factor $a(t)$:

$$\ddot{a} = -\frac{4\pi}{3}(\mu + 3p)a \tag{5.10}$$

$$\dot{a}^2 = \frac{8\pi}{3}\mu a^2 - k \tag{5.11}$$

where $\dot{}$ stands for d/dt. Together (5.10) and (5.11) entail the conservation law

$$\dot{\mu} = -3\frac{\dot{a}}{a}(\mu + p) \tag{5.12}$$

Solving for $a(t)$ requires some assumption about the equation of state linking μ and p. Four epoch's need to be distinguished.[27]

Epoch I: $0 \leq t \leq t_i$

In the inflationary model, as well as in the standard hot big bang model, the universe is radiation dominated immediately after the big bang ($t = 0$). But whereas in the standard model this era lasts until decoupling ($t = t_d$), the inflationary model posits that it ends at a time $t_i < t_d$. Radiation dominance amounts to positing the equation of state $p = \mu/3$. For $k = 0$ (the case I will concentrate on), (5.10) through (5.12) imply that $a(t) \sim t^{1/2}$. In keeping with the notation of Ellis and Stoeger (1988), we have for Epoch I, $a(t) = a_i(2H_i)^{1/2}t^{1/2}$, where $a_i =: a(t_i)$ and H_i is the Hubble constant during this era. Applying the procedure explained in section 5.4, we can now calculate the coordinate horizon distance at the end of Epoch I, with the result

$$r_H(t_i) = \frac{2}{a_i(H_i)^{1/2}}(t_i)^{1/2} = \frac{1}{a_i H_i} \tag{5.13}$$

Epoch II: $t_i \leq t \leq t_f$

This is the inflationary era. The universe cools as it expands during Epoch I. According to the inflationary scenario, when the temperature falls low enough, one or more Higgs fields assume non-zero values, resulting in the breaking of the symmetry that unites the strong, weak, and electromagnetic forces. When the temperature drops below the phase transition temperature, the universe enters a metastable state called the *false vacuum*, which has the strange property that the pressure is negative; in fact, the prediction is that $p = -\mu_f$, $\mu_f > 0$. It follows from (5.12) that μ_f is constant during this time. The stress–energy tensor reduces to $T_{ab} = -\mu_f g_{ab}$, which has the form of a cosmological constant term Λg_{ab} with $\Lambda = -\mu_f$. Thus, the GUT mechanism gives rise to an effective negative cosmological constant, which intuitively means a repulsive force that drives rapid expansion.[28] Formally, using $p = -\mu_f =$ constant and $k = 0$ in (5.10) through (5.11) leads to exponential expansion (de Sitter universe). Assuming that $a(t)$ and $\dot{a}(t)$ are continuous at $t = t_i$ (see Hübner and Ehlers 1991), the time dependence of the scale factor during this epoch is $a(t) = a_i \exp[H_i(t - t_i)]$. The contribution to the horizon coordinate distance is

$$\begin{aligned} r_H(t_f - t_i) &= \frac{1}{a_i H_i}(1 - \exp[-H_i(t_f - t_i)]) \\ &= \frac{1}{a_i H_i}(1 - a_i/a_f) \end{aligned} \tag{5.14}$$

where $a_f =: a(t_f)$. So although the volume of the universe expands enormously during the inflationary phase, the horizon coordinate distance and, therefore, the portion of the universe a light signal is able to traverse does not increase as much as in the preinflationary phase (more on this below).

Epoch III: $t_f \leq t \leq t_d$

The effective cosmological constant now disappears, and we return to the kind of radiation dominance of Epoch I. Thus, from the end of inflation to the decoupling time, the scale factor behaves as $a(t) = a_f[2H_f(t - t_f) + 1]^{1/2}$. The contribution to the horizon coordinate distance during this epoch is

$$r_H(t_d - t_f) = \frac{1}{a_f H_f}(a_d/a_f - 1) \tag{5.15}$$

where $a_d =: a(t_d)$.

Epoch IV: $t_d \leq t \leq t_N$

From the time of decoupling to the present time t_N the universe is assumed to be matter dominated, i.e., the pressure terms in equations (5.10) through

(5.12) are negligible. EFE then imply that $a(t) = a_d[\frac{3}{2}H_d(t - t_d) + 1]^{2/3}$. The contribution to the coordinate horizon distance during this epoch is

$$r_H(t_N - t_d) =: r_{vh} = \frac{2}{a_d H_d}\left[(a_N/a_d)^2 - 1\right] \tag{5.16}$$

where $a_N =: a(t_N)$ and r_{vh} stands for the current visual horizon.

5.12 Does inflation solve the horizon problem?

The original inflationary scenario of Guth (1981) suffered from the defect that the phase transition that creates exponential expansion also creates inhomogeneities greater than allowed by current observational limits. The new inflationary scenario of Linde (1982) and Albrecht and Steinhardt (1982) overcame this defect and provided for a "graceful exit" from inflation. Although there is no direct experimental evidence to support the particle theories used in the inflationary model, the model has gained a wide following principally because (its proponents claim) it overcomes a number of shortcomings of the standard model, including not only the horizon problem but also the flatness and monopole problems.[29]

The validity of the claim that inflation solves the horizon problem depends, of course, on what the problem is supposed to be. In section 5.9 it was suggested that the strong form of the problem poses the challenge of presenting a robust explanation of the presently observed uniformity of the universe. The proponents of inflationary cosmology seem to accept this challenge. The model is then open to two related criticisms. First, it can be charged that the fine tuning of initial conditions in the standard big bang model is matched in the inflationary model by a fine tuning of parameters needed to get the model to agree with observations (see Padmanabhan and Seshadri 1987). Second, it can be charged that in at least one natural sense of measure, the set of initial conditions that leads to inflation is of small measure. Thus, Penrose (1986, 1989b) invites us to consider all possible Cauchy data for the present stage of the universe. It would seem that space slices on which the data are homogeneous and isotropic even at large scales are rare as compared with slices with irregular data. We can then use the backward determinism of EFE to trace backwards generic initial data from the current time to find generic initial data for the big bang singularity. Assuming that the inflationary mechanism is effective in smoothing out irregularities, it follows that inflation does not occur in a generic universe following the big bang.[30] (A response to Penrose's argument is to be found in Turner 1987, pp. 238–239; Penrose's rejoinder is in his 1989b, p. 267; see also Raychadhuri and Modak 1988.)

Setting aside these worries for sake of discussion, it still remains to understand how inflation solves the original puzzle about the homogeneity and isotropy of the CMBR. Indeed, the more reflective reader may be puzzled as to why inflation does not make a bad situation even worse. Consider again the behavior of the particles η and η' in Fig. 5.5. The flying apart of η and η' in Minkowski spacetime might reasonably be thought to mimic some aspects of rapid inflation. If so, it would seem that inflation makes it harder rather than easier for η and η' to interact; for although we have seen that it is a truism that once within a particle horizon, always within a particle horizon, the latter stages of η and η' after hyperbolic acceleration begins ($\approx$inflation) cannot causally interact with one another. And as seen in the preceding section, inflation gives rise effectively to a portion of a de Sitter universe whose conformal structure is indicated in Fig. 5.7. Here we see the case of a particle κ which is in ψ's particle horizon but whose later stages cannot causally interact with ψ. In the inflationary scenario the de Sitter expansion lasts only for a finite time and not into the infinite future; as a result the inflationary and postinflationary stages of particles that are on the edges of each other's particle horizons at the beginning of inflation do not lose forever the ability to causally interact, but these particles do lose the ability to interact via signals during inflation. The analysis of Epoch II shows how the inflationary stages of particles at $r = 0$ and $r = 1/a_i H_i$ cease to be able to causally interact by means of causal signals (see Patzelt 1990 for a discussion of this and related matters).

So how does inflation solve the original puzzle? It does not make particle horizons disappear at the present epoch, contrary to what some enthusiastic statements by proponents of inflation might lead the unwary reader to believe (see Padmanabhan and Seshadri 1987). Rather, the virtue of inflation is to make possible a robust explanation of the presently observed features of the CMBR. From our present spacetime location s (Fig. 5.10) we look back to events p and q at the time of decoupling and observe the CMBR. A robust explanation of the observation requires a satisfaction of the large fraction condition (section 5.7); that is, the intersection of the common causal past $J^-(p) \cap J^-(q)$ of p and q with a time slice $t =$ constant (a measure of the data that can affect both p and q) must become a large fraction of the volume of the intersection of $J^-(p)$ (or of $J^-(q)$) with $t =$ constant (a measure of the data that can affect p (or q) as $t \to 0^+$). In computing this ratio the scale factor $a(t)$ divides out, so we can work in terms of coordinate distances. From Fig. 5.10, we see that what we want to compute is the cube of the ratio $2r_{ccp}/r_x$. But $r_x = 2r_H(t_N) - 2r_H(t_d)$, and $2r_{ccp} = 2r_H(t_N) - 2r_x = 4r_H(t_d) - 2r_H(t_N)$. Since $r_H(t_N) = r_H(t_d) + r_{vh}$, $2r_{ccp}/r_x = r_H(t_d)/r_{vh} - 1$. Of course, when $r_{vh} > r_H(t_d)$ the last ratio does not make sense since then there is no common causal past for p and q. From the calculations of section 5.11 and the relations between the constants $a_f H_f = a_d H_d (a_d/a_f)$ and $a_i H_i = a_f H_f (a_i/a_f)$,[31] we have

$$\frac{r_H(t_d)}{r_{vh}} = \frac{1}{2[(a_N/a_d)^{1/2} - 1]}\left(2\frac{a_f}{a_i}\frac{a_f}{a_d} - 2\frac{a_f}{a_d} + 1\right) \tag{5.17}$$

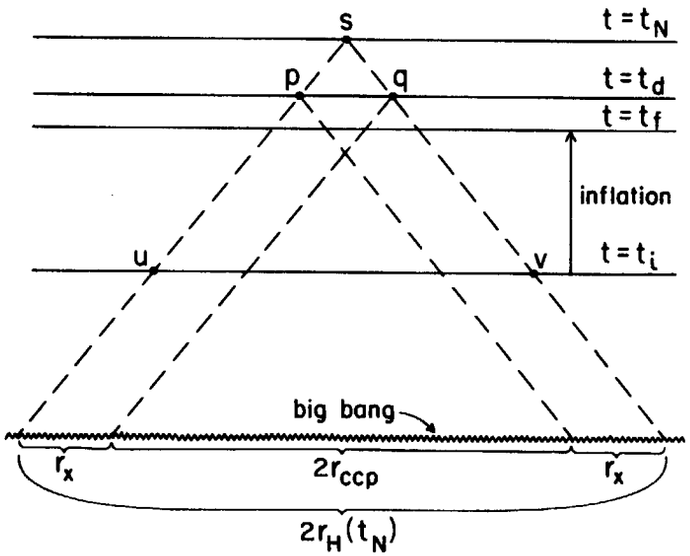

Fig. 5.10 The inflationary model of the big bang in conformal representation

In the standard big bang model, $a_f = a_i$ (no inflation) and (5.17) has a value less than 1, as expected. In the inflationary model a_f/a_i is huge—in most versions it is at least 10^{75}—while a_f/a_d is of the order of 10^{-24}, with the upshot that the numerator on the right-hand side of (5.17) is of the order of 10^{51} (Ellis and Stoeger 1988).

Four remarks about this calculation need to be made, and the first two are closely related. At most, what has been accomplished is a demonstration that the inflationary scenario makes the causal structure of spacetime friendly to a robust dynamical explanation. That the similarity of the CMBR and other conditions at p and q and other similarly situated points is in fact given a robust explanation in terms of the conditions in the common causal past of these points remains to be demonstrated. Only detailed physical calculations can settle this matter (see Ellis and Stoeger 1988 for a discussion of what is involved). The second remark is that for fields that are transmitted at *exactly* the speed of light not even the minimal necessary condition for a common cause explanation has been demonstrated. For the case in question the appropriate ratio is not the one used above; rather, what counts is the limit as $t \to 0^+$ of the ratio of the area of the intersection (t = constant) $\cap\, L^-(p) \cap L^-(q)$ (i.e., the measure of the initial data that can affect the field at both p and q) to the area of the intersection (t = constant) $\cap\, L^-(p)$ (or (t = constant) $\cap\, L^-(q)$) (i.e., the measure of the initial data that can affect p or q as the case may be). This ratio is 0 unless the past light cones of p and q merge, i.e., there are no particle horizons at t_d, which is not the case in the $k = 0$ and $k = -1$ inflationary models.

The third remark concerns a seemingly curious feature of the ratio $r_H(t_d)/r_{vh}$ in the inflationary model. The denominator is the same for both the standard model and the inflationary model, so the difference in the values of the ratio for the two models must lie in the difference in the values assigned to the numerator. In the inflationary model $r_H(t_d) = r_H(t_i) + r_H(t_f - t_i) + r_H(t_d - t_f)$. The third term on the right of the equality is small in comparison with the first two, so the two leading terms must be responsible for the large value of the ratio in (5.17). For strong inflation—a_f/a_i large—$r_H(t_i) \simeq r_H(t_f - t_i)$, as is seen by a glance at (5.13) and (5.14). So the consistency of the inflationary model requires a large contribution to the horizon radius from the initial non-inflationary, radiation-dominated epoch. To put it in teleological terms, the initial epoch in the inflationary model gives a larger contribution than in the standard model because in the inflationary model it is anticipating the later inflation. However, this does not mean that in the inflationary model the horizon problem is already solved by the end of the initial epoch. If we could peer back to $t = t_i$, events such as u and v of Fig. 5.10 have no common causal past.

The fourth remark goes to the ultimate significance of the accomplishment of the inflationary model in making $r_H(t_d)/r_{vh}$ large at the current time. In $k = 0$ and $k = -1$ universes, if enough time elapses, we (or whatever creatures are around at the time) will have a visual horizon that is comparable to or larger than the horizon radius at the decoupling time, in which case the possibility of robust dynamical explanation is lost. Similarly, as advances in neutrino and gravitational wave astronomy allow us to probe further and further backward in time we will eventually be able to "see" events which have no common causal past (see Padmanabhan and Seshardhi 1987, 1988). The proponents of the inflationary model may reply that the fact that their model does not solve every problem does not diminish its accomplishment in solving the problem that was initially posed. The rejoinder is that if problems exactly similar to the original one arise because of advances in technology or even merely waiting around, then the proposed solution to the original problem is fragile and, thus, unsatisfying. So there is an argument to be made that *if* the strong form of the horizon problem is accepted as a genuine problem, then halfway houses such as the inflationary model, which do not get rid of particle horizons, do not provide an adequate solution.

5.13 Conclusion

There is a popular view to the effect that an anomaly for a theoretical framework is recognized as such only after a competing framework has succeeded in resolving the problem (see Lightman and Gingerich 1991). The horizon problem constitutes a counterexample in the sense that particle horizons were perceived to be a problematic feature of standard big bang models long before a satisfactory resolution was reached; indeed, the horizon

problem was one of the motivations for a large number of explorations of alternative cosmological models, no one of which has yet achieved general acceptance. In a deeper sense, however, the popular view is correct. For even now it is far from clear what the horizon problem is and why the standard big bang model should be deemed inadequate because of the presence of particle horizons. Arguably, astrophysicists will have to settle on a model for the big bang before the exact nature of the problem can be defined.

In the meantime, however, some attempt has to be made to delimit the shape of the problem, if only in a rough and preliminary fashion. My attempt here succeeded only in producing a blurry outline. (1) Sure prediction in a generic general relativistic spacetime is impossible. The presence of particle horizons makes the impossible even more so in that it increases the difficulty in making reasonable inductive inferences about the data that may influence the events to be predicted. But this fact does not seem to provide a sufficient basis for consecrating a horizon problem. (2) Since the observed homogeneity and isotropy of the universe within a model with particle horizons involves a coordination between relatively spacelike events which have no common causal past, a conflict with Reichenbach's PCC could be seen. But on closer examination, it was not clear why the pronouncements of the PCC should be heeded for general relativistic cosmological models or, indeed, even what these pronouncements are. (3) Particle horizons are most definitely objectionable from a Machian perspective that demands that all physical effects be tied to ponderable sources, for in generic models with particle horizons there will be source-free electromagnetic and gravitational radiation. But once again it is not clear why such a perspective should be given any currency, especially since from the modern point of view fields enjoy a status as fundamental as particles. (4) The potentially most telling objection to particle horizons starts from the charge that in standard cosmological models such horizons force the use of very special initial conditions. (It was proposed that this insight is the most plausible candidate for the valid core of the PCC as applied to cosmological models with particle horizons.)[32] The objection can be continued in two ways. First, it can be argued that this specialness shows that the prior probability of the standard model is very low. This objection could be overcome by observational evidence indicating that the postulated initial conditions do in fact obtain in our universe. Second, a deeper objection, which would not be overcome by such observational evidence, can be raised to the effect that the standard big bang model lacks robustness and thereby fails to provide a satisfactory explanation of the currently observed uniformity of the CMBR. This complaint is to be taken seriously, but it rests on assumptions about the nature of scientific explanation that require justification. Nor does the currently most popular self-proclaimed solution to the horizon problem provide even a might-makes-right justification. The inflationary model can succeed only by fine-tuning its parameters, and even then, relative to some natural measures on initial conditions, it may also have to fine-tune initial conditions for inflation to work. And the inflationary explanation is fragile in

other ways as well: the robustness of explanation promised by the inflationary scenario may evaporate as we move into the future or probe into the past with new means of detecting events in the early universe.

Perhaps all will be made clear by some new cosmological theory. Or perhaps future generations will conclude that the horizon problem was a tempest in a teapot. Perhaps what is needed is some deflation of the horizon problem rather than inflation of the universe.

Notes

1. Throughout this chapter it is assumed that all spacetimes discussed are temporally orientable (see chapter 6 for a definition of this concept) and that a direction for time has been chosen.

2. At least if the coupled Einstein–matter equations are of the appropriate second-order hyperbolic type; see Wald (1984a, Ch. 10) for a discussion of the exact conditions needed for a well-posed initial value problem.

3. The constraint equations for GTR were stated in chapter 3. For details, see Wald (1984a, p. 259).

4. This example was suggested by Robert Wald.

5. As noted by Geroch (1977). For a review of various senses of prediction in GTR, see Hogarth (1993).

6. The failure of the implications of relativistic theories to count as genuine predictions does not affect the testability or confirmability of these theories.

7. The terms *null cone* and *light cone* are used ambiguously in the literature. Sometimes *null cone* at $p \in M$ is used to denote the object that lies in the tangent space M_p. Other times it is used to denote what I have called the light cone at p.

8. Joshi (1993) says that M, g_{ab} is *causally simple* just in case $J^{\pm}(p)$ are closed for every $p \in M$. This condition is deserving of the name since it simplifies the hierarchy of causality conditions for relativistic spacetimes; see chapter 6.

9. These models are described in more detail later in this section and in section 5.11.

10. For spatially open FRW universes, if there is a particle horizon for our observer at any time, then there is one for all times.

11. The origin of the new time coordinate $\tilde{t}$ can be chosen so that $\tilde{t} = 0$ corresponds to the big bang.

12. This is a little too facile, as Tim Maudlin has kindly told me. One cannot speak of experiments that demonstrate the violation of the Bell inequalities without engaging the measurement problem in QM. This problem takes a particularly nasty form in the relativistic setting. For example, the orthodox treatment of quantum measurement involves a collapse of the state vector. It is not clear whether or not such a collapse can be given a relativistically invariant rendition. Gordon Fleming (1989) argues that such a rendition requires relativizing the quantum state to spacelike hyperplanes. I agree that these are real difficulties whose resolution may require substantial revisions in current physical theory. But it seems to me that the root of these difficulties has to do not with common causes but with much more fundamental issues like the actualization of potentialities. And it is hardly surprising that problems of non-locality and relativistic invariance arise in measurement collapse since the collapse is literally a miracle in the sense of a violation of (what we take to

be) physical laws. Even in paradigms of local classical relativistic field theories (say, electromagnetic field theory in Minkowski spacetime) the introduction of a local miracle (say, the creation of an electrical charge) involves non-local instantaneous effects.

13. For attempts to state a valid kernel of the PCC, see Sober and Barrett (1992) and Arntzenius (1992).

14. Equation (5.5) requires that the electromagnetic potentials $\mathbf{A}$, φ satisfy the *Lorentz gauge* equation $\nabla \cdot \mathbf{A} + \partial\varphi/\partial t = 0$.

15. For homogeneous and isotropic models the effects of charges at the location p in Fig. 5.3 will cancel out so that there is no inconsistency in assuming that incoming fields are zero.

16. In retrospect it seems that the extent to which GTR incorporates Mach type principles is much less than Einstein originally thought; see Raine (1981) for a review of various interpretations of Mach's principle in GTR.

17. Börner is here speaking of the "flatness problem" rather than the horizon problem; but the sentiment is the same for both.

18. This theorem is proved for the context of Minkowski spacetime; but presumably a suitable generalization will hold for curved spacetimes.

19. These correlations cannot be used to send causal signals between the relatively spacelike regions of spacetime.

20. Recall that the *dominant energy condition* requires that for every timelike vector V^a, $T^{ab}V_aV_b \geq 0$ (no negative energy densities) and that $T^{ab}V_a$ is non-spacelike (local energy flow is non-spacelike). The *positive pressure criterion* requires that the sum of the principal pressures of T^{ab} is non-negative.

21. For an account of Bayesian reasoning in science, see Earman (1992).

22. This is a little oversimplified since the uniformity of conditions at earlier times would raise problems for galaxy formation.

23. See chapter 2 for a definition of the Weyl curvature.

24. This matter is discussed further in chapter 6.

25. A semipopular exposition of inflationary cosmology is given in Guth and Steinhardt (1989). A collection of original papers is Abbott (1986). For review articles, see Barrow (1988), Gibbons, Hawking, and Siklos (1983), Blau and Guth (1987), and Turner (1987).

26. Virtually all the standard treatments of horizons and inflation rely on the FRW models. This may seem paradoxical since these models are homogeneous and isotropic whereas inflationary scenarios are supposed to encompass non-homogeneous and non-isotropic initial conditions. The explanation is expediency—calculating the behavior of the scale factor is most easily done in the FRW models. It has to be hoped that conclusions will not be qualitatively different in more complicated models.

27. The analysis of the four epochs given here is essentially a recapitulation of the beautiful analysis of Ellis and Stoeger (1988).

28. The Higgs field is a scalar field φ. The associated stress–energy tensor $\mathcal{T}_{ab} = \nabla_a\varphi\nabla_b\varphi - g_{ab}((1/2)g^{cd}\nabla_c\varphi\nabla_d\varphi - V(\varphi))$, where $V(\varphi)$ is the potential energy. To get $\mathcal{T}_{ab}$ into the form Λg_{ab}, the first term must be zero and the spatial derivatives in the second term must be constant.

29. Following Penrose (1989b) it is useful to distinguish *internal* from *external* problems. The absence of monopoles is to be counted as an internal or self-consistency problem for the inflationary model since it uses GUT theories that generate the problem in the first place. By contrast the homogeneity and flatness problems are external problems. The flatness problem arises from the fact that the density of the universe seems to be very near to the critical density that has to be exceeded if expansion is to be halted.

30. Penrose's objection is a kind of jujitsu move against the inflationary scenario, and it is independent of the details of the GUT mechanisms which are supposed to produce inflation. Mazenko, Unruh, and Wald (1985) have examined in detail the new inflationary scenario and claim that in general it does not lead to exponential expansion; see also Wald (1986).

31. See equations (A4b) and (A4c) of the appendix to Ellis and Stoeger (1988).

32. This is not to say that, at bottom, the aims of the physicists who advocate inflationary cosmology and the philosophers who advocate the PCC are the same. The two groups are superficially joined in common cause in trying to explain correlations between relatively spacelike events. But the driving force behind inflationary cosmology—and several other attempts to solve the horizon problem—is the desire for robust dynamical explanations. The fulfillment of this desire requires not only the existence of a common causal past but also the satisfaction of the large fraction condition (see sections 5.7 and 5.12). By contrast the advocates of the PCC look for an event (or family of events) which lies in the common causal past of the correlated events and which in some (perhaps probabilistic) sense can be said to be the cause of the correlated events. In this search there is no commitment to robustness and no need to satisfy the large fraction condition. But this very lack of commitment produces a tension. For the lack of interfering causes that would prevent the "common cause" event from setting up a correlation between relatively spacelike events, can be viewed as another kind of correlation between distance events that is no less in need of explanation than the original one. I assume that advocates of the PCC would respond that in what we take to be normal background conditions, causal influences propagate without undue interference and that no explanation of the background conditions is called for—perhaps these things are simply part of the very meaning of 'causal propagation' and 'background conditions'. I do not deny that this is so. But I submit that what is going on here has much to do with common sense and very little to do with fundamental physics. Right or wrong, the version of common cause reasoning used in inflationary cosmology reflects one widely shared line of thinking about the structure of good explanations in physics rather than commonsensical reasoning about causes or some philosopher's image of what scientific explanation ought to be like.

6

Time Travel

6.1 Introduction

Over the last few years leading physics journals, such as *Physical Review*, *Physical Review Letters*, *Journal of Mathematical Physics*, and *Classical and Quantum Gravity*, have been publishing articles dealing with time travel and time machines.[1] Why? Have physicists decided to set up in competition with science fiction writers and Hollywood producers? More seriously, does this research cast any light on the sorts of problems and puzzles that have featured in the philosophical literature on time travel?

The last question is not easy to answer. The philosophical literature on time travel is full of sound and fury, but the significance remains opaque. Most of the literature focuses on two matters, backward causation and the paradoxes of time travel.[2] Properly understood, the first is irrelevant to the type of time travel most deserving of serious attention; and the latter, while always good for a chuckle, are a crude and unilluminating means of approaching some delicate and deep issues about the nature of physical possibility. The overarching goal of this chapter is to refocus attention on what I take to be the important unresolved problems about time travel and to use the recent work in physics to sharpen the formulation of these issues.[3]

The plan of this chapter is as follows. Section 6.2 distinguishes two main types of time travel—Wellsian and Gödelian. The Wellsian type is inextricably bound up with backward causation. By contrast, the Gödelian type does not involve backward causation, at least not in the form that arises in Wellsian stories of time travel. This is not to say, however, that Gödelian time travel is unproblematic. This chapter is devoted largely to attempts, first, to get a more accurate fix on what the problems are and, second, to provide an assessment of the different means of dealing with these problems. Section 6.3 provides a brief excursion into the hierarchy of causality conditions on relativistic spacetimes and introduces the concepts needed to assess the problems and prospects of Gödelian time travel. Section 6.4 reviews the known examples of general relativistic cosmological models allowing Gödelian time travel. Since Gödel's (1949a) discovery, it

has been found that closed timelike curves (CTCs) exist in a wide variety of solutions to EFE. This suggests that if classical general relativity theory is to be taken seriously, so must the possibility of Gödelian time travel. Section 6.5 introduces the infamous grandfather paradox of time travel. It is argued that such paradoxes involve both less and more than initially meets the eye. Such paradoxes cannot possibly show that time travel is conceptually or physically impossible. Rather, the parading of the paradoxes is a rather ham-handed way of making the point that local data in spacetimes with CTCs are constrained in unfamiliar ways. The shape and status of these constraints has to be discerned by other means. Section 6.6 poses the problem of the status of the consistency constraints in terms of an apparent incongruence between two concepts of physical possibility that diverge when CTCs are present. Section 6.7 considers various therapies for the time travel malaise caused by this incongruence. The preferred therapy would provide an account of laws of nature on which the consistency constraints entailed by CTCs are themselves laws. I offer an account of laws that holds out the hope of implementing the preferred therapy. This approach is investigated by looking at recent work in physics concerning the nature of consistency constraints for both non-self-interacting systems (section 6.8) and self-interacting systems (section 6.9) in spacetimes with CTCs. Section 6.10 investigates a question that is related to but different from the question of whether time travel is possible; namely, is it possible to build a time machine that will produce CTCs where none existed before? Some concluding remarks are given in section 6.11. An appendix reviews Gödel's attempt to use his solution to EFE to prove the ideality of time.

6.2 Types of time travel; backward causation

Two quite different types of time travel feature in the science fiction and the philosophical literature, though the stories are often so vague that it is hard to tell which is intended (or whether some altogether different mechanism is supposed to be operating). In what I will call the *Wellsian type*[4] the time travel takes place in a garden variety spacetime—say, Newtonian spacetime of classical physics or Minkowski spacetime of special relativistic physics. So the funny business in this kind of time travel does not enter in terms of spatiotemporal structure but in two other places: the structure of the world lines of the time travellers and the causal relations among the events on these world lines. Figure 6.1 illustrates two variants of the Wellsian theme. Figure 6.1a shows the time traveller α_1 cruising along in his time machine. At e_1 he sets the time travel dial to "minus 200 years," throws the switch, and presto he and the machine disappear. Two hundred years prior to e_1 (as measured in Newtonian absolute time or the inertial time of the frame in which α_1 is at rest) a person exactly resembling the time traveller both in terms of physical appearance and mental states pops into existence at e_2. Even if we swallow

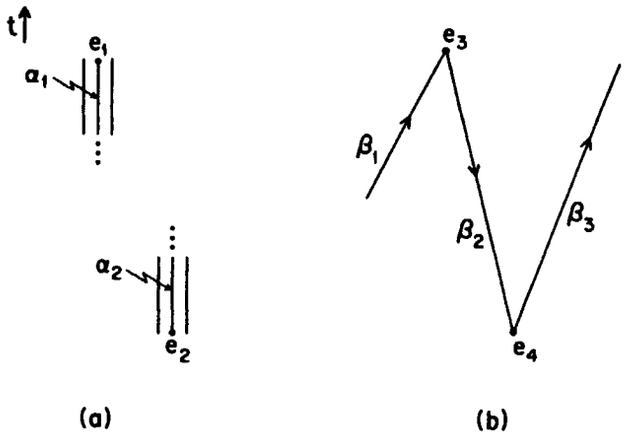

Fig. 6.1 Two forms of Wellsian time travel

these extraordinary occurrences, the description given so far does not justify the appellation of "time travel." That appellation requires that although α_1 is discontinuous with α_2, α_2 is in some appropriate sense a continuation of α_1. Whatever else that sense involves, it seems to require that events on α_1 cause the events on α_2. Thus enters backward causation in which causes are later than their effects.

Figure 6.1b also involves funny world line structure, but now instead of being discontinuous, the world line "bends backwards" on itself, the arrows on the various segments indicating increasing biological time. Of course, as with the previous case, this one also admits an alternative interpretation that involves no time travel. As described in external time, the sequence of events is as follows. At e_4 a pair of middle-aged twins is spontaneously created; the β_3 twin ages in the normal way while his β_2 brother gets progressively younger; meanwhile, a third person β_1 who undergoes normal biological aging and who is the temporal mirror image of β_2 is cruising for a fateful meeting with β_2; when β_1 and β_2 meet at e_3 they annihilate one another. Once again, the preference for the time travel description seems to require a causal significance for the arrows on the world line segments so that, for example, later events on β_1 (as measured in external time) cause earlier events on β_2 (again as measured in external time).

Much of the philosophical literature on Wellsian time travel revolves around the question of whether backward causation is conceptually or physically possible, with the discussion of this question often focusing on the "paradoxes" to which backward causation would give rise. I will not treat these paradoxes here except to say that they bear analogies to the paradoxes of Gödelian time travel that will receive detailed treatment below. But aside from such paradoxes, there is the prior matter of whether the phenomena represented in Fig. 6.1 are physically possible, even when shorn of their time

travel/backward causation interpretations. In Fig. 6.1a, for example, the creation *ex nihilo* at e_2 and the extinction *ad nihilum* at e_1 are at odds with well entrenched conservation principles. Of course, the scenario can be modified so that conservation of mass–energy is respected: at e_1 the time traveler and the time machine dematerialize as before but now their mass is replaced by an equivalent amount of energy, while at e_2 a non-material form of energy is converted into an equivalent amount of ponderable matter. But this emended scenario is much less receptive to a time travel/backward causation reading. For the causal resultants of e_1 can be traced forwards in time in the usual way while the causal antecedents of e_2 can be traced backwards in time, thus weakening the motivation for seeing a causal link going from e_1 to e_2.

At first blush Gödelian time travel would seem to have three advantages over Wellsian time travel. First, on the most straightforward reading of physical possibility—compatibility with accepted laws of physics—Gödelian time travel would seem to count as physically possible, at least as regards the laws of the general theory of relativity (GTR). Second, unlike stories of Wellsian time travel, Gödelian stories are not open to a rereading on which no time travel takes place. And third, no backward causation is involved. On further analysis, however, the first advantage turns out to be something of a mirage since (as discussed below in sections 6.5 through 6.8) Gödelian time travel produces a tension in the naive conception of physical possibility. And the second and third advantages are gained in a manner that could lead one to object that Gödelian time travel so-called is not time travel after all.

To begin the explanation of the claims, I need to say in some detail what is meant by Gödelian time travel. This type of time travel does not involve any funny business with discontinuous world lines or world lines that are "bent backwards" on themselves. Rather, the funny business all derives from the structure of the spacetime which, of course, cannot be Newtonian or Minkowskian. The funny spacetimes contain continuous and even infinitely differentiable timelike curves such that if one traces along such a curve, always moving in the future direction as defined by the globally defined external time orientation, one eventually returns to the very same spacetime location from whence one began. There is no room here for equivocation or alternative descriptions; hence the second advantage. (More cautiously and more precisely, there are some spacetimes admitting Gödelian time travel in the form of closed, future-directed timelike curves, and the curves cannot be unrolled into open curves on which events are repeated over and over ad infinitum—at least such a reinterpretation cannot be made without doing damage to the local topological features of the spacetime; see section 6.3.) As for the third advantage, consider a spacetime M, g_{ab} containing a CTC γ that is instantiated by, say, a massive particle. Pick a point $p \in \gamma$ and choose a small neighborhood $\mathcal{N}$ of p. If $\mathcal{N}$ is chosen wisely, all the causal relations in the restricted spacetime $\mathcal{N}$, $g_{ab|\mathcal{N}}$ will be "normal." So if $q \in \mathcal{N}$ is also on γ and is chronologically later than p, one would judge unequivocally that events

at p cause those at q and not vice versa. But in the encompassing spacetime one might be tempted to say that backward causation is involved since, although q is chronologically later than p, events at q causally influence those at p because γ emerges from N and loops around to rejoin p. But the situation here is quite different from that in Wellsian time travel. In universes with Gödelian time travel it is consistent to assume—and, in fact, is implicitly assumed in standard relativistic treatments—that all causal influences in the form of energy–momentum transfers propagate forward in time with a speed less than or equal to that of light. So in the case at issue, events at q causally influence those at p because q chronologically precedes p and because there is a continuous causal process linking q to p and involving always future-directed causal propagation of energy–momentum. Of course, one could posit that there is another kind of causal influence, not involving energy–momentum transfer, by which events at q affect those at p backwards in time, so that even if the future-directed segment of γ from q to p were to disappear, events at q would still influence those at p. But the point is that Gödelian time travel need not implicate such a backward causal influence.

We are now in a position to see why the second and third advantages have been purchased at a price. One can object that Gödelian time travel does not deliver time travel in the sense wanted since Gödelian time travel so-called implies that there is no time in the usual sense in which to "go back." In Gödel's (1949a) universe, for example, there is no serial time order for events, since every spacetime point p chronologically precedes itself; nor is there a single time slice which would permit one to speak of the Gödel universe at a given time (see section 6.3). I feel that there is a good deal of justice to this complaint. But I also feel that the phenomenon of "time travel" in the Gödel universe and in other general relativistic cosmologies is a worthy object of investigation, whether under the label of "time travel" or under another. The bulk of this chapter is devoted to that investigation.

Before starting on that task, it is worth mentioning for sake of completeness other senses of time travel that appear in the literature. For example, Chapman (1982) and Zemach (1968) devise various scenarios built around the notion of "two times." One interpretation of such schemes would involve the replacement of the usual relativistic conception of spacetime as a four-dimensional manifold equipped with a Lorentz metric of signature $(+++-)$ (three space dimensions plus one time dimension) with a five-dimensional manifold equipped with a metric of signature $(+++--)$ (three space dimensions and two time dimensions). This scheme is worthy of investigation in its own right, but I will confine attention here to standard relativistic spacetimes.

6.3 The causal structure of relativistic spacetimes

There is an infinite hierarchy of causality conditions that can be imposed on relativistic spacetimes (Carter 1971). I will mention only sufficiently many of these conditions to give some flavor of what the hierarchy is like. The review also serves the purpose of introducing the concepts needed for an assessment of Gödelian time travel. To save the reader from having to refer back to previous chapters, I will repeat some of the definitions of key concepts.

The basic presupposition of the causality hierarchy is that of temporal orientability.

(C0) M, g_{ab} is *temporally orientable* iff the null cones of g_{ab} admit of a continuous division into two sets, 'past' and 'future'.[5]

Which set is which is part of the problem of the direction of time, a problem that for present purposes we may assume to have been resolved.

With a choice of temporal orientation in place, we can say that for $p, q \in M$, p *chronologically precedes* q (symbolically, $p \ll q$, or in previous notation $p \in I^-(q)$) just in case there is a smooth future-directed timelike curve from p to q. Similarly, p *causally precedes* q (symbolically, $p < q$, or in previous notation $p \in J^-(q)$) just in case there is a smooth future-directed non-spacelike curve from p to q. It follows without any further restrictions on M, g_{ab} that $\ll$ and $<$ are transitive relations. The first condition of the causality hierarchy says that $\ll$ has the other property we expect of an order relation, namely, irreflexivity.

(C1) M, g_{ab} exhibits *chronology* iff there is no $p \in M$ such that $p \ll p$.

Chronology is, of course, equivalent to saying that the spacetime does not permit Gödelian time travel.

The next condition up the hierarchy is simple causality.

(C2) M, g_{ab} exhibits *simple causality* iff there is no $p \in M$ such that $p < p$.

The next step requires that distinct spacetime points have distinct chronological pasts and futures. Formally,

(C3) M, g_{ab} is *future* (respectively, *past*) *distinguishing* iff for any $p, q \in M$, $I^+(p) = I^+(q) \Rightarrow p = q$ (respectively, $I^-(p) = I^-(q) \Rightarrow p = q$).

An equivalent definition states that M, g_{ab} is future (respectively, past) distinguishing just in case for any $p \in M$ and any neighborhood $\mathcal{N}(p)$, there is a neighborhood $\mathcal{N}'(p) \subseteq \mathcal{N}(p)$ such that no future (respectively, past) directed causal curve from p intersects $\mathcal{N}'(p)$ more than once (Hawking and Ellis 1973, p. 192).

Stronger than both simple causality and past and future distinguishing is the condition of strong causality.

(C4) M, g_{ab} is *strongly causal* iff for any $p \in M$ and any neighborhood $\mathcal{N}(p)$, there is a neighborhood $\mathcal{N}'(p) \subseteq \mathcal{N}(p)$ such that no causal curve intersects $\mathcal{N}'(p)$ more than once.

Intuitively, strong causality not only rules out closed causal curves but also "almost closed" causal curves. Carter (1971) showed that there is a countably infinite hierarchy of conditions lying above (C4) which intuitively rule out "almost almost closed" and "almost almost almost closed" etc. causal curves. Strong causality is still not strong enough to guarantee the existence of a time structure similar to that of familiar Newtonian or Minkowski spacetime, both of which possess a time function. Recall that M, g_{ab} is said to possess a *global time function* just in case there is a differentiable map $t: M \to \mathbb{R}$ such that the gradient of t is a past-directed timelike vector field. This implies that $t(p) < t(q)$ whenever $p \ll q$. The necessary and sufficient condition for such a function is given in the next condition in the hierarchy.

(C5) M, g_{ab} is *stably causal* iff there exits on M a smooth non-vanishing timelike vector field t^a such that M, g'_{ab} satisfies chronology, where $g'_{ab} = g_{ab} - t_a t_b$ and $t_a = g_{ab} t^b$.

Intuitively, stable causality says that it is possible to widen out the null cones of g_{ab} without allowing CTCs to appear. The proof that stable causality implies strong causality uses the fact that a stably causal spacetime possesses a global time function t. For any $p \in M$ and any neighborhood $\mathcal{N}(p)$, a judicious choice of a subneighborhood $\mathcal{N}'(p)$ can be made such that the value of t on any causal curve leaving $\mathcal{N}'(p)$ is greater than its value on the curve when entering $\mathcal{N}'(p)$. Since t must increase along a future-directed causal curve, no such curve can intersect $\mathcal{N}'(p)$ more than once (Wald 1984a, p. 199).

None of the conditions given so far are enough to guarantee that causality in the sense of determinism has a fighting chance on the global scale. That guarantee is provided by a condition already encountered in connection with the discussion of cosmic censorship—namely, global hyperbolicity. The definition of this concept was given in chapter 2. Rather than repeat it here, I will repeat the key fact about it:

(C6) M, g_{ab} is *globally hyperbolic* iff it possesses a Cauchy surface.

Recall that if M, g_{ab} admits one Cauchy surface, then it can be partitioned by them. In fact, a global time function t can be chosen so that each of the level surfaces $t =$ constant is Cauchy.

The conditions (C1) through (C6) form a hierarchy in the sense that each of (C2) through (C6) entails but is not entailed by the one below. There are even stronger causality conditions above (C6), but they will play no role in what follows. The philosophical literature has devoted most of its attention to the ends of the hierarchy, principally to (C0) through (C2) and to (C6),

and has largely neglected (C3) through (C5) and the infinity of other intermediate conditions that have not been enumerated. There are both good and dubious reasons for this selective attention. The intimate connection of (C0) and (C6) respectively to the long-standing philosophical problems of the direction of time and determinism is enough to explain and justify the attention lavished on these conditions. Focusing on (C1) and (C2) to the exclusion of (C3) through (C5) can be motivated by two considerations. First, if one takes seriously the possibility that chronology can be violated, then one must a fortiori take seriously the possibility that everything above can fail. Second, Joshi (1985) showed that (C2) together with a continuity condition, called *causal simplicity*, entail a good bit of the hierarchy above (C2). The condition says that $J^{\pm}(p)$ are closed sets for all $p \in M$. If $J^{+}(p)$ were not closed, there would have to be a situation where $p < q_n$, $n = 1, 2, 3, \ldots$, with $q_n \to q$ but $\neg(p < q)$, i.e., a causal signal can be sent from p to each of the points q_n but not to the limit point q.

Despite these good reasons for the selective focus, I suspect that most of the philosophical attention lavished on (C1) derives from the fascination with the paradoxes of time travel, and that I take to be a dubious motivation. But before taking up this matter in section 6.5, I turn to reasons for taking seriously the possibility of chronology violation.

6.4 Why take Gödelian time travel seriously?

Any relativistic spacetime M, g_{ab} based on a compact M contains CTCs (see chapter 2). Stronger results are derivable for cosmological models M, g_{ab}, T^{ab} of GTR. Tipler (1977a) established that if the cosmological model M, g_{ab}, T^{ab} satisfies EFE without cosmological constant, the weak energy condition (recall that this requires that $T_{ab} V^a V^b \geq 0$ for any timelike V^a), and the generic condition (which requires that every timelike and null geodesic experiences a tidal force at some point in its history), then compactness of M entails that the spacetime is *totally vicious* in that $p \ll p$ for every $p \in M$.

CTCs are not confined to compact spacetimes. In Gödel's (1949a) cosmological model, $M = \mathbb{R}^4$.[6] This example is also important in how it illustrates that the failure of chronology can be *intrinsic* in that chronology cannot be restored by 'unwinding' the CTCs. More precisely, an intrinsic violation of chronology occurs when the CTCs do not result (as in Fig. 6.2b) by making identifications in a chronology-respecting covering spacetime (Fig. 6.2a).

Gödel spacetime is totally vicious. But there are other cosmological models satisfying EFE and the energy conditions where chronology is violated but not viciously. This raises the question of whether it is possible to have a spacetime M, g_{ab} where the chronology-violating set $V \subset M$ is non-empty but so small that it is unnoticeable in the sense of being measure zero. The answer is negative since V is always an open set (see Hawking and Ellis 1973).

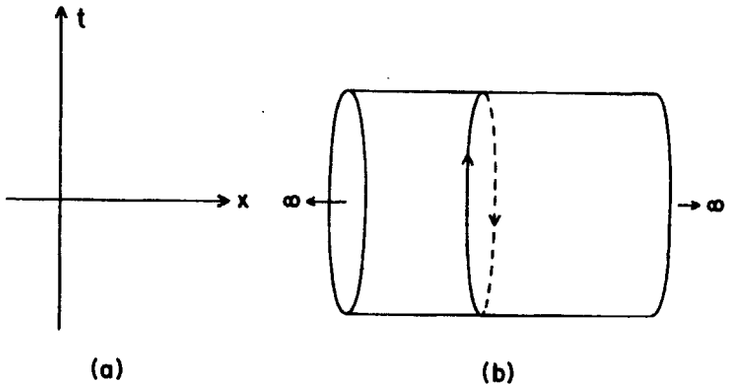

Fig. 6.2 (a) Two-dimensional Minkowski spacetime; (b) rolled-up two-dimensional Minkowski spacetime

In the Gödel universe all CTCs are non-geodesic, necessitating the use of a rocket ship to accomplish the time travel journey. Malament (1985, 1987) provided estimates of the acceleration and the fuel/payload ratio needed to make such a journey. These quantities are so large as to make the journey a practical impossibility. It was on this basis that Gödel (1949b) himself felt justified in ignoring the paradoxes of time travel. Such complacency, however, is not justified. Ozsvath (1967) produced a generalization of Gödel's model that accommodates electromagnetism. De (1969) showed that in such a universe time travellers do not need to use a rocket ship; if they are electrically charged they can use the Lorentz force to travel along CTCs. Even better from the point of view of a lazy would-be time traveller would be a cosmological model with intrinsic chronology violation where some timelike geodesics are closed. An example is provided by the Taub–NUT model which is a vacuum solution to EFE (energy conditions trivially satisfied).

As a result of being totally vicious and simply connected, Gödel spacetime does not contain a single time slice[7] so that one cannot speak of the Gödel universe at a given time. But there are solutions to EFE which are intrinsically chronology violating but which do contain time slices. Indeed, the time slices can themselves be achronal, and thus partial Cauchy surfaces,[8] even though CTCs develop to the future. This raises the question of whether GTR allows for the possibility of building a time machine whose operation in some sense causes the development of CTCs where none existed before. This matter will be taken up in section 6.10.

None of the chronology-violating models discussed so far (with the exception of the trivial example of Fig. 6.2b) are asymptotically flat. But CTCs can occur in such a setting. The Kerr solutions to EFE form a two-parameter family described by the value of the mass M and the angular momentum A. The case where $M^2 > A^2$ is thought to describe the unique final exterior state of a non-charged stationary black hole. In this case

chronology is satisfied. When $M^2 < A^2$ the violation of chronology is totally vicious. Charlton and Clarke (1990) suggest that the latter case could arise if a collapsing rotating star does not dissipate enough angular momentum to form a black hole.

Van Stockum's (1937) solution to EFE, corresponding to a source consisting of an infinite rotating cylinder of dust, contains CTCs. Tipler (1974) suggests that chronology violation may also take place for a finite cylinder source if the rotation rate is great enough.

The Gödel, Ozsvath, Kerr–Newman, and van Stockum models all involve rotating matter. But this is not an essential condition for the appearance of CTCs in models of GTR—recall that the Taub–NUT model is a vacuum solution. Also Morris, Thorne, and Yurtsever (1988) found that generic relative motions of the mouths of traversable "wormholes" (multiply connected surfaces) can produce CTCs, as can generic gravitational redshifts at the wormhole mouths (Frolov and Novikov 1990). However, the maintenance of the traversable wormholes implies the use of exotic matter that violates standard energy conditions. Such violations may or may not be allowed by quantum field theory (see section 6.10).

Gott (1991) found that the relative motion of two infinitely long cosmic strings can give rise to CTCs. This discovery has generated considerable interest and controversy (see Carroll et al. 1992; Deser et al. 1992; Deser 1993; Ori 1991a; 't Hooft 1992). Part of the interest lies in the fact that Gott's solution, unlike the wormhole solutions, does not violate the standard energy conditions of classical GTR. The global structure of Gott's solution has been elucidated by Cutler (1992).

The upshot of this discussion is that since the pioneering work of Gödel some forty years ago, it has been found that CTCs can appear in a wide variety of circumstances described by classical GTR and semiclassical quantum gravity. And more broadly, there are many other known examples of violations of causality principles higher up in the hierarchy. One reaction, which is shared by a vocal if not large segment of the physics community, holds that insofar as these theories are to be taken seriously, the possibility of violations of various conditions in the causality hierarchy, including chronology, must also be taken seriously. This is the attitude of the "Consortium" led by Kip Thorne (see Friedman et al. 1990). Another vocal and influential minority conjectures that GTR has within itself the resources to show that chronology violations can be ignored because, for example, it can be proved that if chronology violations are not present to begin with, they cannot arise from physically reasonable initial data, or because such violations are of "measure zero" in the space of solutions to EFE. This position is championed by Hawking (1992). If such conjectures turn out to be false, one can still take the attitude that in the short run classical GTR needs to be supplemented by principles that rule out violations of the causality hierarchy, and one can hope that in the long run the quantization of gravity will relieve the need for such ad hoc supplementation. Which of these attitudes it is reasonable to

adopt will depend in large measure on whether it is possible to achieve a peaceful coexistence with CTCs. It is to that matter I now turn.

6.5 The paradoxes of time travel

The darling of the philosophical literature on Gödelian time travel is the "grandfather paradox" and its variants. Example: Kurt travels into the past and shoots his grandfather at a time before grandpa became a father, thus preventing Kurt from being born, with the upshot that there is no Kurt to travel into the past to kill his grandfather so that Kurt is born after all and travels into the past . . . (Though the point is obvious, it is nevertheless worth emphasizing that killing one's grandfather is overkill. If initially Kurt was not present in the vicinity of some early segment of his grandfather's world line, then traveling along a trajectory that will take him into that vicinity, even if done with a heart innocent of any murderous intention, is enough to produce an antinomy. This remark will be important for the eventual unraveling of the real significance of the grandfather paradox.)

On one level it is easy to understand the fascination that such paradoxes have exercised—they are cute and their formal elucidation calls for the sorts of apparatus that is the stock-in-trade of philosophy. But at a deeper level there is a meta-puzzle connected with the amount of attention lavished on them. For what could such paradoxes possibly show? (1) Could the grandfather paradox show that Gödelian time travel is not logically or mathematically possible?[9] Certainly not, for we have mathematically consistent models in which CTCs are instantiated by physical processes. (2) Could the grandfather paradox show that Gödelian time travel is not conceptually possible? Perhaps so, but it is not evident what interest such a demonstration would have. The grandfather paradox does bring out a clash between Gödelian time travel and what might be held to be conceptual truths about spatiotemporal/causal order. But in a similar way the twin paradox of special relativity theory reveals a clash between the structure of relativistic spacetimes and what were held to be conceptual truths about time lapse. The special and general theories of relativity have both produced conceptual revolutions. The twin paradox and the grandfather paradox help to emphasize how radical these revolutions are, but they do not show that these revolutions are not sustainable or contain inherent contradictions. (3) Could the grandfather paradox show that Gödelian time travel is not physically possible? No, at least not if "physically possible" means compatibility with EFE and the energy conditions, for we have models which satisfy these laws and which contain CTCs. (4) Could the paradox show that although Gödelian time travel is physically possible it is not physically realistic? This is not even a definite claim until the relevant sense of "physically realistic" is specified. And in the abstract it is not easy to see how the grandfather paradox would support that claim as opposed to the claim that time travel is flatly impossible. Additional factors such as the need

for high accelerations to complete a time travel journey or the instability of Cauchy horizons connected with CTCs (see section 6.10) would seem to be needed to support the charge that Gödelian time travel is physically unrealistic. If anything, such factors tend to mitigate the force of the paradoxes (see sections 6.6, 6.8, and 6.10).

(5) Doesn't the grandfather paradox at least demonstrate that there is a tension between time travel and free will? Of course Kurt cannot succeed in killing his grandfather. But one might demand an explanation of why Kurt doesn't succeed. He had the ability, the opportunity, and (let's assume) the desire. What then *prevented* him from succeeding? Some authors pose this question in the rhetorical mode, suggesting that there is no satisfactory answer so that either time travel or free will must give way. But if the question is intended non-rhetorically, it has an answer of exactly the same form as the answer to analogous questions that arise when no CTCs exist and no time travel is in the offing. Suppose, for instance, that in the time travel scenario Kurt had his young grandfather in the sights of a .30-30 rifle but didn't pull the trigger. The reason the trigger was not pulled is that laws of physics and the relevant circumstances make pulling the trigger impossible at the relevant spacetime location. With CTCs present, *global* Laplacian determinism (which requires a Cauchy surface, as discussed in chapter 3) is inoperable. But *local* determinism makes perfectly good sense. In any spacetime M, g_{ab}, chronology-violating or not, and at any $p \in M$ one can always choose a small enough neighborhood $\mathcal{N}$ of p such that $\mathcal{N}, g_{ab|\mathcal{N}}$ possesses a Cauchy surface Σ with $p \in J^+(\Sigma)$. And the relevant initial data on Σ together with the coupled Einstein–matter equations will uniquely determine the state at p. Taking p to be the location of the fateful event of Kurt's pulling/not pulling the trigger and carrying through the details of the deterministic physics for the case in question shows why Kurt didn't pull the trigger. Of course, one can go on to raise the usual puzzles about free will; namely, granting the validity of what was just said, is there not a way of making room for Kurt to have exercised free will in the sense that he could have done otherwise? At this point all of the well-choreographed moves come into play. There are those (the *incompatibilists*) who will respond with arguments intended to show that determinism implies that Kurt couldn't have done otherwise, and there are others (the *compatibilists*) waiting to respond with equally well-rehearsed counterarguments to show that determinism and free will can coexist in harmony. But all of this has to do with the classic puzzles of determinism and free will and not with CTCs and time travel per se.

(6) Perhaps we have missed something. Suppose that Kurt tries over and over again to kill his grandfather. Of course, each time Kurt fails—sometimes because his desire to pull the trigger evaporates before the opportune moment, sometimes because although his murderous desire remains unabated his hand cramps before he can pull the trigger, sometimes because although he pulls the trigger the gun misfires, sometimes because although the gun fires the bullet is deflected, etc. In each instance we can give a deterministic

explanation of the failure. But the obtainment of all the initial conditions that result in the accumulated failures may seem to involve a coincidence that is monstrously improbable (see Horwich 1989). Here we have reached a real issue but one which is not easy to tackle.

A first clarificatory step can be taken by recognizing that the improbability issue can be formulated using inanimate objects. (Consider, for example, the behavior of the macroscopic objects in my study as I write: a radiator is radiating heat, a light bulb is radiating electromagnetic waves, etc. If the world lines of these objects are CTCs, it would seem to require an improbable conspiracy to return these objects to their current states, as required by the completion of the time loop.) Since free will is a murky and controversial concept, it is best to set it aside in initial efforts at divining the implications of the grandfather paradox. After some progress has been made it may then be possible to draw some consequences for free will. As a second step we need to formalize the intuition of improbability. One method would be to define a measure on the space of solutions to EFE and to try to show that the solutions corresponding to some kinds of time travel (those involving the functional equivalent of Kurt trying over and over again to kill his grandfather) have negligible or flatly zero measure. Even if such a demonstration is forthcoming, we still have to face the question: So what? (After all, some types of space travel will be measure zero, but this hardly shows that the concept of space travel is suspect.) The answer will depend crucially on the justification for and significance of the measure. This matter will receive some attention in section 6.8. But for the moment I want to note that the impression of improbability in connection with time travel stories may not be self-reenforcing. In the above example the judgment of the improbability of the failure of Kurt's repeated attempts to kill his grandfather was made relative to our (presumably chronology respecting) world; but perhaps from the perspective of the time travel world itself there is no improbability. By way of analogy, suppose that the actual world is governed by all the familiar laws of classical relativistic physics *save for* Maxwell's laws of electromagnetism. If we peered into another world which was nomologically accessible from our world but which was governed by Maxwell's laws we would see things that from our perspective are improbable ("measure zero") coincidences. We would find, for example, that the electric and magnetic fields on a time slice cannot be freely specified but must satisfy a set of constraints; and we would find that once these constraints are satisfied at any moment they are thereafter maintained for all time (see chapter 5). Amazing! But, of course, from the perspective of the new world there is no improbability at all; indeed, just the opposite is true since the "amazing coincidences" are consequences of the laws of that world. That this analogy may be opposite to the case of time travel will be taken up in sections 6.6 and 6.7.

What then remains of the grandfather paradox? The paradox does point to a seemingly awkward feature of spacetimes that contain CTCs: local data are constrained in a way that is not present in spacetimes with a more normal

causal structure. But the forms in which the paradox has been considered in the philosophical literature are of little help in getting an accurate gauge of the shape and extent of the constraints. And by itself the paradox is of no help at all in assessing the status of these consistency constraints.

6.6 Consistency constraints

The laws of special and general relativistic physics that will be considered here are all local in the following twofold sense. First, they deal with physical situations that are characterized by local geometric object fields O (e.g., scalar, vector, tensor fields) on a manifold M. Second, the laws governing these fields are in the form of local ordinary or local partial differential equations. The result is a *global-to-local property*: if M, g_{ab}, O satisfies the laws and $U \subseteq M$ is an open neighborhood, then $U, g_{ab}|_N, O|_U$ also satisfies the laws. (This property holds whether or not CTCs are present.) Thus, it would seem at first blush that the question of whether some local state of affairs is physically possible can be answered by focusing exclusively on what is happening locally and ignoring what is happening elsewhere.

In Minkowski spacetime and in general relativistic spacetimes with nice causality properties we typically have the reverse *local-to-global property*: any local solution can be extended to a global solution.[10] Consider, for example, the source-free wave equation for a massless scalar field Φ: $g^{ab}\nabla_a\nabla_b\Phi =: \Box\Phi = 0$, where ∇_a is the derivative operator associated with g_{ab}. On Minkowski spacetime ($M = \mathbb{R}^4$ and $g_{ab} = \eta_{ab}$ (Minkowski metric)), any C^∞ solution on an open $U \subset \mathbb{R}^4$ can be extended to a full solution on $\mathbb{R}^4$. But obviously this local-to-global property fails for the chronology-violating spacetime of Fig. 6.2b. Figure 6.3a shows a local solution with a single pencil of rays traversing U. This solution is obviously globally inconsistent since the light rays from U will trip around the spacetime and reintersect U.

The point is straightforward, but some attempts to elaborate it make it sound mysterious. Thus, consider the presentation of the Consortium:

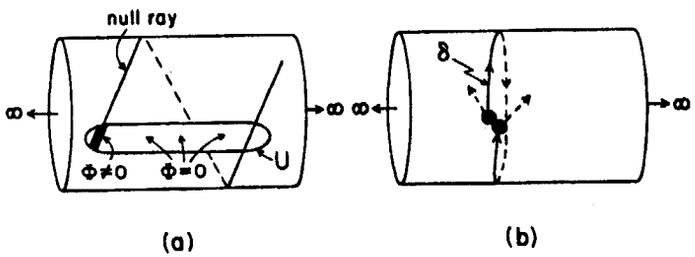

Fig. 6.3 (a) Light rays and (b) billiard balls in rolled-up Minkowski spacetime

> The only type of causality violation the authors would find unacceptable is that embodied in the science-fiction concept of going backward in time and killing one's younger self ("changing the past" [grandfather paradox]). Some years ago one of us (Novikov) briefly considered the possibility that CTC's might exist and argued that they cannot entail this type of causality violation: Events on a CTC are already guaranteed to be self-consistent, Novikov argued; they influence each other around the closed curve in a self-adjusted, cyclical, self-consistent way. The other authors have recently arrived at the same viewpoint.
>
> We shall embody this viewpoint in a *principle of self-consistency*, which states that *the only solutions to the laws of physics that can occur locally in the real universe are those which are globally self-consistent*. This principle allows one to build a local solution to the equations of physics only if that local solution can be extended to be part of a (not necessarily) unique global solution, which is well defined throughout the nonsingular regions of spacetime. (Friedman et al. 1990 pp. 1916–1917)

The first part of the quotation seems to invoke either a notion of preestablished harmony or else a guiding hand that prevents the formation of an inconsistent scenario. But once such connotations are removed, the "principle of self-consistency" (PSC) threatens to deflate into a truism. Here is the Consortium's comment:

> That the principle of self-consistency is not totally tautological becomes clear when one considers the following alternative: The laws of physics might permit CTC's; and when CTC's occur, they might trigger new kinds of local physics, which we have not previously met. ... The principle of self-consistency is intended to rule out such behavior. It insists that local physics is governed by the same types of physical laws as we deal with in the absence of CTC's. ... If one is inclined from the outset to ignore or discount the possibility of new physics, then one will regard self-consistency as a trivial principle. (ibid., p. 1917)

What the Consortium means by discounting the possibility of "new physics" is, for example, ignoring the possibility that propagation around a CTC can lead to multivalued fields, calling for new types of laws that tolerate such multivaluedness. I too will ignore this possibility. But I will argue shortly that CTCs may call for "new physics" in another sense. For the moment, however, all I want to insist upon is that taking the PSC at face value seems to force a distinction between two senses of physical possibility.

In keeping with the global-to-local property introduced above, we can say again more pedantically what was already said informally: a local situation is *physically possible*$_1$ just in case it is a local solution of the laws. But the PSC—which says that the only solutions that can occur locally are globally self-consistent—seems to require a more demanding and relativized sense of physical possibility; namely, a local situation is *physically possible*$_2$ in a spacetime M, g_{ab} just in case the solution can be extended to a solution of the laws on all of M, g_{ab}. If the conditions a local solution must fulfill in order to be extendible to a global solution are called the *consistency*

constraints, one can roughly paraphrase physical possibility$_2$ as physical possibility$_1$ plus satisfaction of the consistency constraints.[11]

This distinction might be regarded as desirable for its ability to let one have one's cake and eat it too. On the one hand, we have the intuition that it is physically possible to construct and launch a rocket probe in any direction we like with any velocity less than that of light. This intuition is captured by physical possibility$_1$. But on the other hand it is not possible to realize all of the physically possible$_1$ initial conditions for such a device in spacetimes with certain kinds of CTCs since the traverse of a CTC may lead the probe to interfere with itself in an inconsistent way. (Or so it would seem; but see the discussion of section 6.8.) This impossibility is captured by the failure of physical possibility$_2$.

On reflection, however, having one's cake and eating it too is, as usual, too good to be true. Thus, one might maintain with some justice that to be physically impossible simply is to be incompatible with the laws of physics—as codified in the definition of physical possibility$_1$. So as it stands the notion of physical impossibility$_2$ seems misnamed when it does not reduce to physical impossibility$_1$, as it apparently does not when CTCs are present. To come at the point from a slightly different angle, let us reconsider the grandfather paradox. It was suggested in the preceding section that Kurt's failure to carry out his murderous intentions could be explained in the usual way—by reference to conditions that obtained before the crucial event and the (locally) deterministic evolution of these conditions. But while not incorrect, such an explanation deflects attention from a doubly puzzling aspect of spacetimes with CTCs. First, it may not even be possible for Kurt to set out on his murderous journey, much less to carry out his intentions. And second, the ultimate root of this impossibility does not lie in prior contingent conditions since there are no such conditions that can be realized in the spacetime at issue and which eventuate in the commencement of the journey. The ultimate root of this impossibility taps the fact that (as we are supposing) there is no consistent way to continue Kurt's murderous journey in the spacetime. But, one might complain, to call this impossibility "physical impossibility$_2$" is to give a label that is not backed by any explanatory power; for given the way the story has been told so far the local conditions corresponding to the commencement of Kurt's journey are compatible with all the postulated laws. In such a complaint the reader will detect a way of trying to scratch the residual itch of the grandfather paradox.

The itch must be dealt with once and for all. I see three main treatments, the first two of which promise permanent cures while the third denies the existence of the ailment.

6.7 Therapies for time travel malaise

(T1) This treatment aims at resolving the tension between physical possibility$_1$ and physical possibility$_2$ by getting rid of the latter. The leading idea is to

argue that GTR shows that, strictly speaking, the notion of consistency constraints is incoherent. For example, looking for consistency constraints in a spacetime M, g_{ab} for a scalar field obeying $\Box\Phi = 0$ makes sense if Φ is treated as a test field on a fixed spacetime background. But (the argument continues) this is contrary to both the letter and spirit of GTR. For Φ will contribute to the total energy–momentum—the usual prescription being that $T_{ab}(\Phi) = \nabla_a\Phi\nabla_b\Phi - 1/2 g_{ab}\nabla_c\Phi\nabla^c\Phi$—that generates the gravitational field cum metric. And (one could conjecture) if Φ and Φ' are interestingly different (say, they differ by more than an additive constant), then the metrics g_{ab} and g'_{ab} solving EFE for the corresponding $T_{ab}(\Phi)$ and $T'_{ab}(\Phi')$ will be different (i.e., non-isometric). This therapy is radical in that if it succeeds it does so by the draconian measure of equating the physically possible$_1$ local states with the actual states. This is intuitively unsatisfying. If we restrict attention to Φ's such that $T_{ab}(\Phi)$ is small in comparison with the total T_{ab} and the spacetime is stable under small perturbations of T_{ab}, then Φ can to good approximation be treated as a test field. Questions of stability will be examined in section 6.10, but in the meanwhile I will assume that they can be set aside. One could also object that (T1) is inapplicable in cases where there are CTCs and where the laws entail that the spacetime structure is non-dynamical (in the sense of not responding to the matter content) and that a variety of physically possible$_1$ states can be realized on a given local region. However, the strength of this objection is hard to grasp since the laws in question would have to be rather different from those of our world, at least if something akin to GTR is true. And recall that the success of GTR is the main reason for taking Gödelian time travel seriously.

(T2) The second treatment strategy is to naturalize physical possibility$_2$. The idea is, first, to insist that physical possibility (relative to a world) just is the compatibility with the laws (of that world) and, second, to go on to argue that physical possibility$_2$ can be brought into the fold by showing that in chronology-violating environments the consistency constraints of physical possibility$_2$ have law status. Thus, (T2) insists that, contrary to the Consortium's explanation of the PSC, there is a sense in which CTCs do call forth "new physics."[12]

(T2) can take two forms. (a) The naturalization of physical possibility$_2$ would amount to a reduction to physical possibility$_1$, understood as consistency with the local laws of physics, if the consistency constraints/new laws were purely local so that, even in the chronology-violating environments, what is physically possible locally is exactly what is compatible with the (now augmented) local laws of physics. (b) Unfortunately, the reduction of (a) can be expected in only very special cases. In general, the consistency constraints may have to refer to the global structure of spacetime. In these latter cases, insofar as (T2) is correct, the concept of physical possibility$_1$ must be understood to mean consistency of the local situation with all the laws, local and non-local. The patient who demands a purely local explanation of the difference between local conditions that are physically possible and those which are not will continue to itch.

(T3) If the first two therapies fail, the discomfort the patient feels can be classified as psychosomatic. The therapist can urge that the patient is getting overexcited about nothing or at least about nothing to do specifically with time travel; for global features of spacetime other than CTCs can also impose constraints on initial data. For example, particle horizons in standard big bang cosmologies prevent the implementation of the Sommerfeld radiation condition which says that no source-free electromagnetic radiation comes in from infinity (see chapter 5). Here the patient may brighten for a moment only to relapse into melancholy upon further reflection. For the constraints entailed by the particle horizons are of quite a different character than those entailed by the typical chronology-violating environment; the former, unlike the latter, do not conflict with the local-to-global property and thus do not drive a wedge between physical possibility$_1$ and physical possibility$_2$. And in any case the choice of the particle horizons example is not apt for therapeutic purposes since these horizons are widely thought to be so problematic as to call for new physics involving cosmic inflation or other non-standard scenarios.[13] Clearly this line of therapy opens up a number of issues that require careful investigation; but such an investigation is beyond the scope of this book.

My working hypothesis favors (T2). This is not because I think that (T2) will succeed; indeed, I am somewhat pessimistic about the prospects of success. Nevertheless, making (T2) the focus of attention seems justified on several grounds: the success of (T2) would provide the most satisfying resolution of the nagging worries about time travel, while its failure would have significant negative implications for time travel; whether it succeeds or fails, (T2) provides an illuminating perspective from which to read recent work on the physics of time travel; and finally, (T2) forces us to confront issues about the nature of the concept of physical law in chronology-violating spacetimes, issues which most of the literature on time travel conveniently manages to avoid.

It is well to note that my working hypothesis is incompatible with some analyses of laws. For example, Carroll (1994) rejects the idea that laws supervene on occurrent facts,[14] and adopts two principles which have the effect that laws of the actual world $W_@$ are transportable *as laws* to other possible worlds which are nomologically accessible from $W_@$. The first principle says that "if P is physically possible and Q is a law, then Q would (still) be a law if P were the case" (p. 59). The second says that "if P is physically possible and Q is not a law, then Q would (still) not be a law if P were the case" (p. 59). Let P say that spacetime has the structure of Gödel spacetime or some other spacetime with CTCs. And let us agree that P is physically possible because it is compatible with the laws of $W_@$ (which for sake of discussion we may take to be the laws of classical general relativistic physics). It follows from Carroll's principles that if spacetime were Gödelian, the laws of $W_@$ would still be laws and also that these would be the only laws that would obtain.

I will not attempt to argue here for the supervenience of laws on occurrent

facts but will simply assume it. In exploring my working hypothesis, I will rely on an account of laws that can be traced back to John Stuart Mill (1843); its modern form is due to Frank Ramsey (1928/29) and David Lewis (1973). The gist of the Mill–Ramsey–Lewis (MRL) account is that a law for a logically possible world W is an axiom or theorem of the best overall deductive system for W (or what is common to the systems that tie for the best). A deductive system for W is a deductively closed, axiomatizable, set of (non-modal) sentences, each of which is true in W. Deductive systems are ranked by how well they achieve a compromise between strength or information content on the one hand and simplicity on the other. Simplicity is a notoriously vague and slippery notion, but the hope is that, regardless of how the details are settled, there will be for the actual world a clearly best system or at least a non-trivial common core to the systems that tie for best. If not, the MRL theorist is prepared to admit that there are no laws for our world.

The MRL account of laws is naturalistic (all that exists is spacetime and its contents), actualistic (there is only one actual world); and empiricistic (a world is a totality of occurrent facts; there are no irreducible modal facts). In addition, I would claim that this account fits nicely with the actual methodology used by scientists in search of laws. The reader should be warned, however, that it is far from being universally accepted among philosophers of science; see, for example, van Fraassen (1989) and my response in Earman (1993). I will not attempt any defense of the MRL account here. If you like, the ability to illuminate the problems of time travel can be regarded as a test case for the MRL account.

Suppose for sake of discussion that the actual world has a spacetime without CTCs; perhaps, for example, its global features are described more or less by one of the FRW big bang models. And suppose that the MRL laws of this world are just the things dubbed laws in textbooks on relativistic physics, no more, no less. Now consider some other logically possible world whose spacetime contains CTCs. But so as not to waste time on possibilities that are too far removed from actuality, let us agree to restrict attention to worlds that are nomologically accessible from the actual world in that the laws of the actual world, taken as non-modal propositions, are all true of these worlds. Nevertheless, one cannot safely assume that the MRL laws of our world "govern" these time travel worlds in the sense that the set of laws of our world coincides with the set of laws of time travel worlds.

One possibility is that the MRL laws of a time travel world W consist of the MRL laws of this world *plus* the consistency constraints on the test fields in question. If so, we have a naturalization of physical possibility$_2$, though it would remain to be seen whether the naturalization takes the preferred (a) form or the less desirable (b) form. Additionally, time travel would have implications for free will. In cases where an action is determined by the laws plus contingent initial conditions, compatibilists and incompatibilists split on whether the actor could be said to have the power to do otherwise

and whether the action is free. But all parties to the free will debate agree that if an action is precluded by the laws alone without the help of contingent boundary or initial conditions, then the action is not in any interesting sense open to the agent. Thus, if the possibility under discussion pans out, there are various actions that, from a compatibilist perspective at least, we are free to perform in this world that we are not free to perform in various time travel scenarios.

Other possibilities also beg for consideration. For instance, it could turn out that although (by construction) the MRL laws of this world are all *true of* a time travel world W, they are not all *laws of* W, except in a very tenuous sense. I will argue that this possibility is realized in cases where the consistency constraints are so severe as to supplant the laws of this world. In such cases the time travel involved is arguably such a remote possibility that it loses much of its interest. But note that since the consistency constraints are still subsumed under the laws of the time travel world, we retain the desirable feature that physical possibility$_2$ is naturalized.

Finally, these remarks point to the intriguing possibility that purely local observations can give clues to the global structure of spacetime without the help of a supplementary "cosmological principle"; namely, local observations may reveal the absence of consistency constraints that would have to obtain if we inhabited certain kinds of chronology-violating spacetimes.

What is needed as a first step in coming to grips with these matters is a study of the nature of consistency conditions on test fields that arise for various chronology-violating spacetimes. The recent physics literature has made some progress on the project. In the next two sections I will report on some of the results for self-interacting and non-self-interacting fields. On the basis of these results I will advance some tentative conclusions to the series of questions posed above concerning physical possibility and laws in chronology-violating worlds.

6.8 Non-self-interacting test fields

When considering chronology-violating spacetimes, the simplest regime to study mathematically is the case of a non-self-interacting field, e.g., solutions to the source-free scalar wave equation $\Box\Phi = 0$. Of course, the grandfather paradox and the related paradoxes of time travel that have been discussed in the philosophical literature typically rely on self-interacting systems. Even so, we shall see that non-trivial consistency conditions can emerge in the non-self-interacting regime. But on the way to illustrating that point it is worth emphasizing the complementary point that in small enough regions of some chronology-violating spacetimes the consistency constraints for non-self-interacting fields do not make themselves felt so that local observations in such regions will not reveal the presence of CTCs.

Following Yurtsever (1990), call an open $U \subseteq M$ *causally regular* for the

spacetime M, g_{ab} with respect to the scalar wave equation just in case for every C^∞ solution Φ of $\Box\Phi = 0$ on U, there is a C^∞ extension to all of M, i.e., there is a C^∞ $\tilde{\Phi}$ on M such that $\Box\tilde{\Phi} = 0$ and $\tilde{\Phi}|_U = \Phi$. M, g_{ab} can be said to be *causally benign* with respect to the scalar wave equation just in case for every $p \in M$ and every open neighborhood U of p there is a subneighborhood $U' \subset U$ which is causally regular.

The two-cylinder of Fig. 6.2b is causally benign with respect to the scalar wave equation. The following remarks, while not constituting a proof of this fact, give an indication of why it holds in the optical limit. In that limit Φ waves propagate at the speed of light (i.e., along null trajectories). At any point on the cylinder a small enough neighborhood can be chosen such that any null geodesic leaving this neighborhood in either the future or past direction never returns. Consider then any solution on this neighborhood. To extend this local solution to a global one, simply propagate the solution out of the base neighborhood along null geodesics. If the propagated field does not reach a point $q \in M$, set $\Phi(q) = 0$. If two null geodesics from the base neighborhood cross at q, obtain $\Phi(q)$ by adding the propagated fields.

Consider next the toroidal spacetime $\mathcal{T}_{(1,r)}$ obtained from two-dimensional Minkowski spacetime by identifying the points (x, t) and (x', t') when $x = x' \bmod 1$ and $t' = t \bmod r$, where $r > 0$ is a real number. For r rational, $\mathcal{T}_{(1,r)}$ is benign with respect to the scalar wave equation, as shown by Yurtsever (1990). Through any point on $\mathcal{T}_{(1,r)}$ there is a time slice that lifts to many $t =$ constant surfaces in the Minkowski covering spacetime $\mathbb{R}^2, \eta_{ab}$. Consider one such surface, say, $t = 0$. Any solution $\Phi(x, t)$ on $\mathcal{T}_{(1,r)}$ induces on $t = 0$ initial data $\Phi_0(x) =: \Phi(x, 0)$ and $\dot{\Phi}_0(x) =: (d/dt)\Phi(x, t)|_{t=0}$. By considering solutions on $\mathbb{R}^2, \eta_{ab}$ of the wave equation that develop from this initial data it follows that both Φ_0 and $\dot{\Phi}_0$ must be periodic with periods 1 and r. Further, $\dot{\Phi}_0$ must satisfy the integral constraint

$$\int_{(t=0)} \dot{\Phi}_0(s)\, ds = 0.$$

When r is rational we can choose a small enough neighborhood of any point on $\mathcal{T}_{(1,r)}$ such that arbitrary initial data Φ_0 and $\dot{\Phi}_0$ can be extended so as to meet the periodicity and integral constraints.

Friedman et al. (1990) argue that the benignity property with respect to the scalar wave equation also holds for a class of chronology-violating spacetimes that are asymptotically flat and globally Minkowskian except for a single wormhole that is threaded by CTCs. In some of these spacetimes there is a partial Cauchy surface Σ such that chronology violations lie entirely to the future of Σ. It is argued that the formation of CTCs places no consistency constraints on initial data specified on Σ.

In all the examples considered so far the chronology violations are non-intrinsic in that they result from making identifications of points in

a chronology-preserving covering spacetime. Unfortunately, because of the non-trivial mathematics involved, almost nothing is known about the benignity properties of spacetimes with intrinsic chronology violations. I conjecture that Gödel spacetime is benign with respect to the scalar wave equation. The conjecture is based on the fact that null geodesics in Gödel spacetime are not only open but are never almost closed, i.e., for any point of Gödel spacetime and any open neighborhood of that point there is a subneighborhood such that every inextendible null geodesic starting in the neighborhood eventually leaves and never returns. Thus, it may be possible to generalize to Gödel spacetime the construction indicated above for extending local solutions to global solutions on the two-cylinder.[15]

If correct, this conjecture would cast new light on a puzzling feature of Gödel's (1949b) own attitude towards the grandfather paradox in Gödel spacetime. Basically his attitude was one of "why worry" since the fuel requirements on a rocket needed to realize a time travel journey in Gödel spacetime are so demanding as to be impossible to meet by any practical scheme. But consistency constraints are constraints whether or not they can be tested by practical means. So it would seem that whatever puzzles arise with respect to the status of such constraints are unresolved by appeal to practical considerations. However, the above conjecture, if correct, suggests that for non-self-interacting systems the consistency constraints in Gödel spacetime are much milder than one might have thought. Similarly, Gödel's remarks can be interpreted as suggesting that the constraints will also be mild for self-interacting systems. Some further information on this matter is obtained in section 6.8.[16]

It should be emphasized at this juncture that in spite of the connotations of the name, benignity does not necessarily imply physics as usual. For it does not imply that there are no non-trivial consistency constraints nor that the constraints cannot be detected locally. Benignity implies only that the constraints cannot be felt in sufficiently small neighborhoods, but this is compatible with their being felt in regions of a size that we typically observe.

To give an example of a non-benign spacetime we can return to $\mathcal{T}_{(1,r)}$ and choose r to be an irrational number. Now the periodicity constraints on the initial data cannot be satisfied except for Φ_0 and $\dot{\Phi}_0$ constant. The integral constraint then requires that $\dot{\Phi}_0 = 0$, with the upshot that the only solutions allowed are $\Phi =$ constant everywhere. No local solution that allows the tiniest variation in Φ can be extended to a global solution.

I turn now to the question of the status of the consistency constraints for chronology-violating spacetimes. By way of introduction, I note the assertion of the Consortium that a time traveler "who went through a wormhole [and thus around a CTC] and tried to change the past would be *prevented by physical law* from making the change" (Friedman et al. 1990, p. 1928; italics added). One way of interpreting this assertion is in line with my working hypothesis that the consistency constraints entailed by the presence of CTCs in a nomologically accessible chronology-violating world are laws of that

world. This position is arguably endorsed by the MRL account of laws. For it is plausible that in each of the above examples the consistency constraints would appear as axioms or theorems of (each of) the best overall true theories of the world in question.

For the reasons discussed in sections 6.6 and 6.7, such a result is devoutly to be desired. But to play the devil's advocate for a moment, one might charge that the result is an artifact of the examples chosen. Each of these examples involves a spacetime with a very high degree of symmetry, and it is this symmetry one suspects of being responsible for the relative simplicity of the consistency constraints. If this suspicion is correct, then the consistency constraints that obtain in less symmetric spacetimes may be so complicated that they will not appear as axioms or theorems of any theory that achieves a good compromise between strength and simplicity. Due to the technical difficulties involved in solving for the consistency constraints in non-symmetric spacetimes, both the devil and his advocate may go blue in the face if they each hold their breath while waiting for a confirmation of their suspicions. If the devil's advocate should prove to be correct, the proponent of the naturalization could still find comfort if the cases where the naturalization fails could be deemed to be very remote possibilities.

An illustration of how time travel can be justly deemed to involve very remote possibilities occurs when the chronology violating world W is so far from actuality that, although the laws of the actual world are *true of* W, they are not *laws of* W except in a very attenuated sense. In the case of $\mathcal{T}_{(1,r)}$ with r irrational we saw that Φ = constant are the only allowed solutions of the scalar wave equation. I take it that $\Box\Phi = 0$ will not appear as an axiom of a best theory of the $\mathcal{T}_{(1,r)}$ world so that the scalar wave equation is demoted from fundamental law status. Presumably, however, Φ = constant will appear as an axiom in any best theory. Of course, this axiom entails that $\Box\Phi = 0$; but it also entails any number of other differential equations that are incompatible with one another and with the scalar wave equation when Φ is not constant. Thus, in the $\mathcal{T}_{(1,r)}$ world the scalar wave equation and its rivals have much the same status as "All unicorns are red," "All unicorns are blue," etc. in a world where it is a law that there are no unicorns. In this sense $\Box\Phi = 0$ has been supplanted as a law. This still is a case where physical possibility$_2$ is naturalized by reduction in the strong sense to physical possibility$_1$ since the consistency constraint is stated in purely local terms. But the more remote the possibility, the less interesting the reduction. And in this case the possibility can be deemed to be very remote since the relation of nomological accessibility has become non-symmetric—by construction the toroidal world in question is nomologically accessible from the actual world, but the converse is not true because the toroidal law Φ = constant is violated here.

One might expect that such supplantation will take place in any world with a spacetime structure that is not benign. What I take to be a counterexample to this expectation is provided by the four-dimensional

toroidal spacetime $\mathcal{T}_{(1,1,1,1)}$ obtained from four-dimensional Minkowski spacetime by identifying the points (x, y, z, t) and (x', y', z', t') just in case the corresponding coordinates are equal mod 1. This spacetime is not benign with respect to the scalar wave equation. But solutions are not constrained to be constant; in fact, the allowed solutions form an infinite-dimensional subspace of the space of all solutions (Yurtsever 1990). The consistency constraint imposes a high frequency cutoff on plane wave solutions propagating in certain null directions. This constraint by my reckoning is simple and clean enough to count as an MRL law, but it supplements rather than supplants $\Box\Phi = 0$.

It is possible in principle to verify by means of local observations that we do not inhabit a non-causally regular region of some non-benign spacetimes. And if we indulge in an admittedly dangerous inductive extrapolation on the above examples, we can conclude that we do not in fact inhabit a non-regular region. For it follows from the (source free) Maxwell equations—which we may assume are laws of our world—that the components of the electromagnetic field obey the scalar wave equation. But the electromagnetic fields in our portion of the universe do not satisfy the restrictions which (if the induction is to be believed) are characteristic of non-benign spacetimes. Further experience with other examples of non-benign spacetimes may serve to strengthen or to refute this inference.

It is also possible to use local observations to rule out as models for our world certain benign chronology-violating spacetimes, but only on the assumption that we have looked at large enough neighborhoods to reveal the consistency constraints indicative of these spacetimes. It is not easy to see how we would come by a justification for this enabling assumption.

6.9 Self-interacting test fields

I begin with a reminder of two lessons from previous sections. First, the grandfather paradox is in the first instance a way of pointing to the presence of consistency constraints on local physics in a chronology-violating spacetime. And second, while the usual discussion of the grandfather paradox assumes a self-interacting system, we have found that non-trivial and indeed very strong constraints can arise even for non-interacting systems. Of course, one would expect that the constraints for self-interacting systems will be even more severe. To test this expectation one could carry out an analysis that parallels that of section 6.8 by considering a test field obeying an equation such as $\Box\Phi = k\Phi^3$ (where k is a constant with sign chosen so as to lead to positive energy density), which implies that solutions do not superpose. Results are not available in the literature. What are available are results for the simpler and more artificial case of perfectly elastic billiard balls.

One assumes that between collisions the center of mass of such a ball traces out a timelike geodesic in the background spacetime and that in a

collision the laws of elastic impact are obeyed. Under these assumptions the initial trajectory δ of the (two-dimensional) billiard ball in the spacetime of Fig. 6.3b leads to an inconsistent time development. The ball trips around the cylinder and participates in a grazing collision with its younger self, knocking its former self from the trajectory that brought about the collision (grandfather paradox for billiard balls). We saw in the previous section that the cylinder can be benign for the scalar wave equation. But obviously the corresponding property fails for billiard balls: for any point x on the cylinder it is not the case that there is a sufficiently small neighborhood $\mathcal{N}(x)$ such that any timelike geodesic segment on $\mathcal{N}$, representing the initial trajectory of the ball, can be extended to a globally consistent trajectory. Nor are the forbidden initial conditions of measure zero in any natural measure.

For a single sufficiently small billiard ball, Gödel spacetime is benign, not just because all timelike geodesics are open but also because they are not almost closed. And it seems safe to conjecture that Gödel spacetime is benign with respect to any finite system of small billiard balls since it seems implausible that collisions among a finite collection can be arranged so as to achieve the sustained acceleration needed to instantiate a closed or almost closed timelike curve.

Echeverria et al. (1991) have studied the behavior of billiard balls in two types of wormhole spacetimes that violate chronology. The first type is called the *twin paradox spacetime* because the relative motions of the wormhole mouths give rise to a differential aging effect which in turn leads to CTCs since the wormhole can be threaded by future-directed timelike curves. This spacetime contains a partial Cauchy surface Σ. Chronology is violated only to the future of Σ, indeed to the future of $H^+(\Sigma)$. The other type of wormhole spacetime is called the *eternal time machine spacetime* because CTCs that traverse the wormhole can reach arbitrarily far into the past. There are no partial Cauchy surfaces in this arena; but because of asymptotic flatness the notion of past null infinity $\mathscr{I}^-$ is well defined, and initial data can be posed on $\mathscr{I}^-$.

For initial conditions (specified on Σ for the twin paradox spacetime or on $\mathscr{I}^-$ for the eternal time travel spacetime) that would send the billiard ball into the wormhole, one might expect to find that strong consistency constraints are needed to avoid the grandfather paradox. But when Echeverria et al. searched for forbidden initial conditions, they were unable to find any. Thus, it is plausible, but not proven, that for each initial state of the billiard ball, specified in the non-chronology-violating region of the spacetime, there exists a globally consistent extension. Mikheeva and Novikov (1992) have argued that a similar conclusion holds for an inelastic billiard ball.

Surprisingly, what Echeverria et al. did find was that each of many initial trajectories had a countably infinite number of consistent extensions. The consistency problem and the phenomenon of multiple extensions is illustrated in Fig. 6.4. In Fig. 6.4a we have yet another instance of the grandfather paradox. The initial trajectory ζ, if prolonged without interruption, takes the

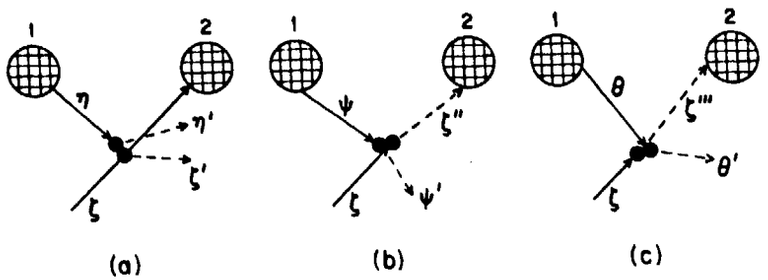

Fig. 6.4 (a) Self-inconsistent and (b, c) self-consistent billiard ball motions in a wormhole spacetime (after Echeverria et al. 1991)

billiard ball into mouth 2 of the wormhole. The ball emerges from mouth 1 along η, collides with its younger self, converting ζ into ζ' and preventing itself from entering the wormhole. Figures 6.4b and 6.4c show how ζ admits of self-consistent extensions. In Fig. 6.4b the ball suffers a grazing collision which deflects it along trajectory ζ''. It then reemerges from mouth 1 along ψ and suffers a glancing blow from its younger self and is deflected along ψ'. The reader can provide the interpretation of Fig. 6.4c. The demonstration of the existence of initial conditions that admit an infinite multiplicity of consistent extensions involves the consideration of trajectories that make multiple wormhole traversals; the details are too complicated to be considered here.

These fascinating findings on the multiplicity of extensions are relevant to the question of whether it is possible to operate a time machine; this matter will be taken up in section 6.10. Of more direct relevance to present concerns are the findings about consistency constraints for self-interacting systems. The results of Echeverria et al. (1991) indicate that in the twin paradox spacetime, for instance, the *non*-chronology-violating portion of the spacetime is benign with respect to all billiard ball trajectories, including those dangerous trajectories that take the ball into situations where the grandfather paradox might be expected. But the chronology-violating region of this spacetime is most certainly *not* benign with respect to billiard ball motions. Perhaps it is a feature of non-benign spacetimes that the failure of benignity only shows up in the chronology-violating region, but one example does not give much confidence.

The study of more complicated self-interacting systems quickly becomes intractable at the level of fundamental physics. What one has to deal with is a coupled set of equations describing the self- and cross-interactions of particles and fields. Deriving properties of solutions of such a set of equations for chronology-violating spacetimes is beyond present capabilities. Instead, one studies the behavior of "devices" whose behavior is analyzed on the macrolevel. The presumption is that if these devices were analyzed into fundamental constituents and if the field equations and equations of motion

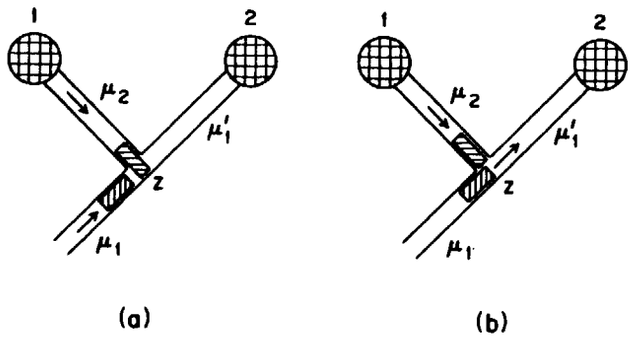

Fig. 6.5 (a) Self-inconsistent and (b) self-consistent motions for Novikov's piston device (after Novikov 1992)

for these constituents were solved for some relevant range of initial and boundary conditions, then the solutions would display the behavior characteristic of the type of device in question. From this point of view, a perfectly elastic billiard ball might be considered to be one of the simplest devices.

A slightly more complicated device consists of a rigid Y-shaped tube and a piston which moves frictionlessly in the tube. Imagine that the branches at the top of the Y are hooked to the mouths of the wormhole in the twin paradox spacetime. Because of the constraints the tube puts on the motion of the piston, it is not true that every initial motion of the piston up the bottom of the Y has a self-consistent extension. Figure 6.5a shows an initial motion that takes the piston up along the sections μ_1 and μ_1', into the wormhole mouth 2, through the wormhole, out of mouth 1 at an earlier time, down the section μ_2 to the junction $\mathcal{Z}$ just in time to block its younger self from entering the μ_1' section.

Novikov (1992) argues that self-consistent solutions are possible if the device is made slightly more realistic by allowing the older and younger versions of the piston to experience friction as they rub against one another. In Fig. 6.5b the piston starts with the same initial velocity as in Fig. 6.5. But when the piston tries to pass through the junction $\mathcal{Z}$ it is slowed down by rubbing against its time-traveling self. This slowing down means that when the piston traverses the wormhole it will not emerge at an earlier enough time to block the junction but only to slow its younger self down. Novikov gives a semiquantitative argument to show that, with self-friction present, for any initial motion of the piston that gives rise to an inconsistent/grandfather paradox evolution as in Fig. 6.5a, there is a self-consistent extension as in Fig. 6.5b. For initial trajectories that have the time-traveling piston arriving at the junction well before its younger self reaches that point, there is also arguably a self-consistent continuation. Thus, it is plausible that there are no non-trivial consistency constraints on the initial motion of the piston up the μ_1 section of the tube.

The work of Wheeler and Feynman (1949) and Clarke (1977) suggests that the absence of consistency constraints or at least the benignity of the constraints can be demonstrated generally for a class of devices for which the evolution is a continuous function of the parameters describing the initial conditions and the self-interaction, the idea being that fixed point theorems of topology can be invoked to yield the existence of a consistent evolution. However, Maudlin (1990) showed that if the topology of the parameter space is complicated enough, a fixed point/self-consistent solution may not exist for some initial conditions. And one would suppose that in the general case the problem of deciding whether the relevant state space topology admits a fixed point theorem is as difficult as solving directly for the consistency constraints.

One might expect that with sufficiently complicated devices there may be no (or only rare) initial conditions that admit a self-consistent continuation in a chronology-violating environment that allows the device to follow CTCs. Consider Novikov's (1992) device consisting of a radio transmitter, which sends out a directed beam; a receiver, which listens for a signal; and a bomb. The device is programmed to detonate the bomb if and only if it detects a signal of a strength that would be experienced by being, say, 30 m from the device's transmitter. A self-consistent traverse of the wormhole of the twin paradox spacetime is possible if the device undergoes inelastic collisions; for then such a collision between the older and younger versions can produce a change in orientation of the transmitter such that the younger self does not receive the signal from its older self and, consequently, no explosion takes place. But one can think of any number of epicycles that do not admit of any obvious self-consistent solution. For example, as Novikov himself suggests, the device could be equipped with gyrostabilizers that maintain the direction of the radio beam.

It is all too easy to get caught up in the fascinating details of such devices and thereby to lose sight of the implications for what I take to be the important issues about time travel. As a way of stepping back, let me reiterate the point that came up in connection with the investigations of Echeverria et al. (1991) of billiard ball motions in chronology-violating spacetimes. The absence of consistency constraints or the benignity of these constraints with respect to the initial conditions of a device, as specified in the non-chronology-violating portion of the spacetime, does not establish the absence of consistency constraints or their benignity simpliciter. For example, assuming some self-friction of the piston of the Novikov cylinder–piston device, there may be no non-trivial consistency constraints on the initial motion of the piston up the bottom of the Y. But there most certainly are constraints on the motion in the chronology-violating portion of the spacetime, and these constraints are not benign. Consider any spacetime neighborhood that includes the junction $\mathcal{Z}$ at a time when the piston is passing $\mathcal{Z}$. Passing from μ_1 to μ_1' with a speed v without rubbing against a piston coming down μ_2 is a physically possible local state for every v. But for some values of v there is no self-consistent extension. Thus, contrary to what some commentators have suggested, the

recent work on the physics of time travel does not dissolve the paradoxes of time travel. Whatever exactly these paradoxes are, they rest on the existence of consistency constraints entailed by field equations/laws of motion in the presence of CTCs. Showing that these constraints are trivial would effectively dissolve the paradoxes. But all of the recent work affirms the non-triviality of such constraints, which are more or less severe depending on the case.

Does what we have learned about self-interacting systems give reason for optimism about my working hypothesis; namely, that insofar as a chronology-violating world admits a set of MRL laws for test fields, those laws will subsume the consistency constraints forced out by the presence of CTCs? One might see a basis for pessimism deriving from the fact that in the wormhole spacetimes the constraints that obtain in different regions are different (e.g., no constraints on the initial conditions for the billiard ball in the non-chronology-violating region of the twin paradox spacetime but non-trivial constraints in the chronology-violating region). Since we want laws of nature to be "universal" in the sense that they hold good for every region of spacetime, it might seem that the wormhole spacetimes dash the hope that the consistency constraints will have a lawlike status. But the hope is not to be extinguished so easily. To be "universal," the constraint must be put in a general form; namely, for any region R, constraint $C(R)$ obtains iff _____, where the blank is filled with conditions formulated in terms of suitably general predicates. The blank will need to be filled not only with features of R but also with features of the relation of R to the rest of spacetime. So if the consistency constraints have law status, then the laws of a chronology-violating wormhole spacetime cannot all be local. But that was only to be expected. The real concern is the one that already surfaced in section 6.7; namely, that as the spacetime gets more and more complicated, the conditions that go into the blank may have to become so complicated that the consistency constraints will not qualify as MRL laws. Remember, however, that this concern is mitigated if the chronology-violating worlds in question can be deemed to lie in the outer reaches of the space of worlds nomologically accessible from the actual world. Here I think that intuition pumping is useless until we have more concrete examples to serve as an anchor.

6.10 Can we build and operate a time machine?

The question that forms the title of this section is not equivalent to the question of whether time travel is possible, at least not if 'time machine' is understood in the strong sense of a device that manufactures or produces CTCs. A sufficiently powerful rocket engine that allowed a person to trace out a CTC in Gödel spacetime might be counted as a time machine in the weak sense. But clearly there is no time machine in the strong sense operating in this context, since CTCs exist everywhere and everywhen.[17]

In trying to characterize a time machine we face something of a conundrum. To make sure that the CTCs are due to the operation of the time machine we could stipulate that there is a time slice Σ, corresponding to a time before the machine is turned on, such that there are no CTCs in $J^-(\Sigma)$. Furthermore, by going to a covering spacetime, if necessary, we can guarantee that Σ is achronal and, thus, a partial Cauchy surface. But then by construction only time travel to the future of Σ is allowed, which immediately eliminates the kind of time travel envisioned in the typical time travel story of the science fiction literature. I do not see any easy way out of this conundrum, and for present purposes I will assume that such a Σ exists.

Next, one would like a condition which says that only local manipulation of matter and energy is involved in the operation of the machine. Requiring that the spacetime be asymptotically flat would be one approach—intuitively, the gravitational field of the machine falls off at large distances. But on the one hand this requirement precludes many plausible cosmologies; and on the other hand it is not evident that a condition on the space structure guarantees that something funny with the causal structure was not already in progress before the machine was switched on.

It is even more delicate to pin down what it means to say that switching on the time machine produces CTCs. Programming the time machine corresponds to setting initial conditions on the partial Cauchy surface Σ. These conditions together with the coupled Einstein–matter equations determine a unique evolution for the portion of $J^+(\Sigma)$ contained in $D^+(\Sigma)$. But $D^+(\Sigma)$ is globally hyperbolic and therefore contains no CTCs. Moreover, the future boundary $H^+(\Sigma)$ of $D^+(\Sigma)$ is always achronal. So the notion that the initial conditions on Σ are responsible for the formation of CTCs cannot be cashed in terms of causal determinism. Perhaps this notion can be captured by the requirement that some of the null geodesic generators of $H^+(\Sigma)$ are closed or almost closed, indicating that CTCs are on the verge of forming. Perhaps it should also be required that any appropriate extension of the spacetime across $H^+(\Sigma)$ contains CTCs. Here an appropriate extension might variously be taken to be one that is sufficiently smooth, that preserves various symmetry properties of $D^+(\Sigma)$, . . .

One need not be too fussy about the sufficient conditions for the operation of a time machine if the goal is to prove negative results, for then one need only fix on some precise necessary conditions. An example of such a negative result was obtained recently by Hawking (1992). In concert with the above discussion he assumes the existence of a partial Cauchy surface Σ such that $H^+(\Sigma)$ separates the portion of spacetime with CTCs from the portion without CTCs. In Hawking's terminology, $H^+(\Sigma)$ is a *chronology horizon*. If all the past-directed null geodesic generators of $H^+(\Sigma)$ enter and remain within a compact set, then $H^+(\Sigma)$ is said to be *compactly generated*.

Theorem (Hawking). Let M, g_{ab}, T_{ab} be a cosmological model satisfying Einstein's field equations (with or without cosmological constant). Suppose

that M, g_{ab} admits a partial Cauchy surface Σ and that T_{ab} satisfies the weak energy condition.[18] If further $H^+(\Sigma)$ is non-empty and compactly generated, then (a) Σ cannot be non-compact, and (b) whatever the topology of Σ, matter cannot cross $H^+(\Sigma)$.

How effective is this formal result as an argument against time machines? At best, part (b) of the theorem shows that the operator of the time machine cannot himself sample the fruits of his labor by crossing over the chronology horizon to the region of spacetime containing CTCs. Part (a) has real bite if it can be read as saying that in a spatially open universe a time machine cannot operate without violating the weak energy condition.[19] However, this reading assumes that when CTCs are manufactured by a time machine, the chronology horizon $H^+(\Sigma)$ must be compactly generated. Physically this requirement says that the generators of $H^+(\Sigma)$ do not emerge from a curvature singularity, nor do they "come from infinity." These prohibitions might seem well motivated by the idea that if the appearance of CTCs is to be attributed to operation of a time machine, then they must result from the manipulation of matter in a finite region of space. But it seems to me this motivation is better served by requiring that (i) $H^+(\Sigma)$ is *compactly causally generated* in the sense that the topological closure of $I^-(H^+(\Sigma)) \cap \Sigma$—which is the portion of Σ from which events can influence $H^+(\Sigma)$—is compact, and (ii) all appropriate extensions across $H^+(\Sigma)$ contain CTCs. It is possible in principle for $H^+(\Sigma)$ to be compactly causally generated although not compactly generated in Hawking's sense; in particular, the former does not preclude that some generators of $H^+(\Sigma)$ emerge from curvature singularities. The would-be time machine operator may well be willing to create singularities in order to satisfy his client's desire to experience the thrill of time travel. Thus, it seems to me that an effective chronology protection theorem would have to substitute the condition of compactly causally generated for Hawking's condition of compactly generated.[20] But when the substitution is made, Hawking's proof technique no longer works. How likely is it that some other technique will yield an effective chronology protection theorem? The answer depends on one's attitude towards the cosmic censorship hypothesis, for such a theorem would constitute a proof of an important piece of cosmic censorship. Given how hard it has been to prove censorship theorems (see chapter 3), one should not expect effective chronology protection theorems to sprout like mushrooms.

A different approach to showing that the laws of physics are unfriendly to the enterprise of building a time machine would be to try to show that the operation of the machine involves physical instabilities. More specifically, in terms of the setting suggested above, one would try to show that a chronology horizon is necessarily unstable. This approach links back to the problem of the behavior of test fields on chronology-violating spacetimes; indeed, the stability property can be explicated in terms of the existence of extensions of solutions of the test field. To return to the example of a scalar field Φ obeying

the wave equation $\Box\Phi = 0$, consider arbitrary initial data on the partial Cauchy surface Σ (the values on Σ of Φ and its normal derivative to Σ) of finite energy.[21] Each such data set determines a unique solution in the region $D^+(\Sigma)$. $H^+(\Sigma)$ can then be said to be *completely stable* for Φ iff every such solution has a smooth extension across $H^+(\Sigma)$.[22] $H^+(\Sigma)$ can be said to be *generically stable* for Φ iff a generic solution has a smooth extension across $H^+(\Sigma)$.[23]

It has been argued by Morris et al. (1988) and by Friedman and Morris (1991a, b) that some asymptotically flat wormhole spacetimes with CTCs have stable chronology horizons. However, these examples violate the weak energy condition of classical GTR. This follows from the results of Tipler (1976, 1977a) showing that the weak energy condition prevents the kind of topology change which occurs with the development of wormholes. Such violations are tolerated in quantum field theory. But the maintenance of traversable wormholes leading to CTCs also requires the violation of averaged or integrated versions of the energy conditions. What remains unresolved is the extent to which quantum field theory tolerates violations of the averaged energy conditions. Wald and Yurtsever (1991) demonstrated the satisfaction of the averaged null energy condition[24] in two-dimensional curved spacetimes; but they also showed that this condition can fail in four-dimensional spacetimes. The proponents of chronology protection can hope that quantum fields can never violate the averaged energy conditions in such a way that permits wormhole-based CTCs.

An example of a spacetime that contains a partial Cauchy surface Σ with CTCs to the future of Σ and where the chronology horizon $H^+(\Sigma)$ is generically unstable for Φ is Misner's two-dimensional version of Taub–NUT spacetime (see Fig. 2.3 and Hawking and Ellis 1973, pp. 170–178). Here the chronology horizon is not only compactly generated but is itself compact; indeed, it is generated by a smoothly closed null geodesic. Each time the tangent vector of this geodesic is transported parallel to itself around the loop it is expanded by a factor of e^h, $h > 0$, indicating a blueshift. Now consider a generic high frequency wave packet solution to $\Box\Phi = 0$ "propagating to the right." As it nears $H^+(\Sigma)$ it experiences a blueshift each time it makes a circuit around the universe, and as an infinite number of circuits are needed to reach $H^+(\Sigma)$, the blueshift diverges. This is already indicative of an instability, but to demonstrate that the divergent blueshift involves the kind of instability that prevents an extension across $H^+(\Sigma)$ it has to be checked that the local energy density of the wave packet diverges as $H^+(\Sigma)$ is approached.[25] That is in fact the case in this example. One could then reason that when Φ is not treated merely as a test field but as a source for the gravitational field, spacetime singularities will develop on $H^+(\Sigma)$ thereby stopping the spacetime evolution and preventing the formation of CTCs that would otherwise have formed beyond the chronology horizon (see Morris et al. 1988).

However, the classical instability of chronology horizons is certainly not

a generally effective mechanism for ensuring chronology protection. Hawking (1992) showed that among compactly generated chronology horizons, a non-negligible subset of the horizons are classically stable. Even in cases where classical instability obtains, one can also wonder how this instability undermines the feasibility of operating a time machine. The worst case of instability would be complete instability with respect to Φ, i.e., no solution of $\Box\Phi = 0$ other than $\Phi = 0$ is extendible across $H^+(\Sigma)$. Then insofar as the time machine involves non-zero values of Φ, it cannot succeed. The next worst case would involve instability that is not complete but is so generic that only a set of solutions of "measure zero" admit extensions across $H^+(\Sigma)$. Here one could argue that if the time machine operator chose the parameter setting at random (with respect to the preferred measure on initial conditions), she would have a zero probability of hitting on a setting that would lead to successful operation of the machine. This would not be a proof of the impossibility of time travel via a time machine but only a demonstration that initiating the journey requires luck. Perhaps some stronger conclusion can be derived, but I do not see how. Measure zero arguments are commonly assumed to have a good deal of force, but it is hardly ever explained why.

In sum, it seems fair to say that at present no mechanisms from classical GTR have been shown to be effective enforcers of chronology protection. If classical GTR does not offer chronology protection, then perhaps it can be found in quantum effects. In particular, the quantum instability of chronology horizons is currently under intensive investigation (see Boulware 1992; Hawking 1992; Kim and Thorne 1991; Klinkhammer 1992). It seems that in the wormhole spacetimes with CTCs, the expectation value of the (renormalized) stress–energy tensor of a quantum field diverges as the chronology horizon is approached, with the divergence being stronger for cases of a compactly generated horizon than for a non-compactly generated horizon. In the semiclassical approach to quantum gravity, the expectation value of the stress–energy tensor is fed back into EFE to determine the effects on the spacetime geometry. Whether or not the divergence of the stress–energy tensor on the chronology horizon produces an alteration of the spacetime sufficient to prevent the formation of CTCs is still controversial. Apparently in Gott spacetime the stress–energy tensor for a scalar field remains regular near the chronology horizon and therefore cannot prevent the formation of CTCs (see Boulware 1992).

Now suppose for the sake of discussion that neither classical GTR nor quantum mechanics prevents the construction of a time machine. The main puzzle about its operation is not the grandfather paradox but something quite different. The implicit assumption in the science fiction literature is that when the time machine is switched on, some definite scenario will unfold, as determined by the settings on the machine. But the still imperfectly understood physics of time travel hints at something quite at variance with these expectations. In the first place, there may be different extensions of the spacetime across the chronology horizon $H^+(\Sigma)$. In the second place, even

when the spacetime extension is chosen and treated as a fixed background for test fields, billiard balls, and other devices, the equations which govern these systems may permit a multiplicity, perhaps even an infinite multiplicity, of extensions across $H^+(\Sigma)$ and into the time travel region. The point is not simply that one does not know the upshot of turning on the time machine but rather that the upshot is radically underdetermined on the ontological level. And thus the new puzzle: How does the universe choose among these ontologically distinct possibilities? Of course, it is unfair to demand a mechanism for making the choice if "mechanism" implies determinism, for that is what is expressly ruled out in this situation. But it is equally unsatisfactory to respond with nothing more than the formula that "It is just a matter of chance which option will unfold." If "just a matter of chance" is to be more than an incantation or a recapitulation of the puzzle, then "chance" must mean something like objective propensity. But the physics of classical GTR provides no basis for saying that there are objective probabilities of, say, .7 and .3 respectively for the scenarios of Fig. 6.4b and Fig. 6.4c. Here quantum mechanics may come to the rescue of time travel by showing that for any initial quantum state describing the motion of the billiard ball before it enters the region of CTCs, there is a well-defined probability for each of the subsequent classically consistent extensions.[26]

However, when CTCs are present it is to be expected that the time evolution will not be unitary (see Friedman et al. 1992; Goldwirth et al. 1993; Politzer 1992, 1994). The loss of unitarity is not necessarily fatal to a viable quantum description; perhaps, for example, the path integral or sum-over-histories approach will provide a means for consistently assigning probabilities to measurement outcomes (see Friedman et al. 1992). Such an approach, even if consistent, certainly exhibits some disturbing features, e.g., the probability for the outcome of a measurement made in the pre-time travel region can depend upon whether CTCs form in the future (see Friedman et al. 1992; Politzer 1994). Clearly, how CTCs mesh or fail to mesh with quantum mechanics will be an exciting area of investigation for some time to come.

6.11 Conclusion

If nothing else, I hope this chapter has made it clear that progress on understanding the problems and prospects of time travel is not going to come from the sorts of contemplations of the grandfather paradox typical of past philosophical writings. Using modal logic to symbolize the paradox, armchair reflections on the concept of causation and the like are not going to yield new insights. The grandfather paradox is simply a way of pointing to the fact that if the familiar laws of classical relativistic physics are supposed to hold true in a chronology-violating spacetime, then consistency constraints emerge. The first step to understanding these constraints is to define their shape and

content. This involves solving problems in physics, not armchair philosophical reflections.

But philosophy can help in understanding the status of the consistency constraints. Indeed, the existence of consistency constraints is a strong hint—but nevertheless a hint that most of the literature on time travel has managed to ignore—that it is naive to expect the laws of a time travel world which is nomologically accessible from our would will be identical with the laws of our world. I explored this matter under the assumption that laws of nature are to be constructed following the analysis of Mill–Ramsey–Lewis. In some time travel worlds it is plausible that the MRL laws include the consistency constraints; in these cases the grandfather paradox has a satisfying resolution. In other cases the status of the consistency constraints remains obscure; in these cases the grandfather paradox leaves a residual itch. Those who wish to scratch the itch further may want to explore other analyses of laws. Indeed, time travel would seem to provide a good testing ground for competing analyses of laws.

I do not see any prospect for proving that time travel is impossible in any interesting sense. It may be, however, that it is not physically possible to operate a time machine that manufactures CTCs. But if so, no proof of this impossibility has emerged in classical GTR. The prospects for getting such a proof are intimately tied to the fate of cosmic censorship. If the operation of the time machine is feasible there emerges a new puzzle: a setting of the parameters on the time machine may correspond to many different scenarios in the time travel region. The problem here is not that the operation of the time machine is unpredictable or calls into play an element of indeterminacy; rather, the problem lies in providing an objective content to the notion of chance in this setting. Quantum mechanics is, of course, the place to look for such content. But standard quantum mechanics is hard to reconcile with CTCs. And it would be a little surprising and more than a little disturbing if Gödelian time machines, which seemed to be characterizable in purely classical relativistic terms, turned out to be inherently quantum mechanical. Is nothing safe from the clutches of the awful quantum?

Appendix: Gödel on the ideality of time

As the reader of this chapter will no doubt have gathered, I think that too much of the philosophical literature on time travel has been devoted to Gödel spacetime. It would be healthier if attention were directed to other solutions to EFE which allow for time travel and which do not exhibit one or other of the peculiarities of Gödel spacetime (e.g., time travel in the Gödel universe requires a fantastically powerful rocket engine whereas in other solutions time travel may be accomplished without the help of rocketry). By contrast, there has been a relative neglect of the philosophical moral Gödel (1949a) himself

wanted to draw from his solution to EFE.[27] I will try to explain why the neglect has been benign. Some explanation is called for, if for no other reason than because a deeply held conviction of someone of Gödel's stature deserves serious consideration.

Unless otherwise indicated, page references are to Gödel (1949b). The dialectic of his argument goes as follows. He began with the idea that if STR were true, then time would be ideal. He wrote:

> Change becomes possible only through the lapse of time. The existence of an objective lapse of time, however, means (or, at least, is equivalent to the fact) that reality consists of an infinity of layers of "now" which come into existence successively. But, if simultaneity is something relative [as is implied by STR] . . . reality cannot be split up into such layers in an objectively determined way. Each observer has his own set of "nows," and none of these various systems of layers can claim the prerogative of representing the objective lapse of time. (p. 558)

In a footnote, Gödel acknowledges that a possible response to this argument is that it shows "that time is something relative, which does not exclude that it is something objective; whereas the idealists maintain that it is something merely imagined" (p. 558, fn 5). Gödel's rejoinder is emphatic.

> A lapse of time, however, which is not a lapse in some definite way seems to me as absurd as a coloured object which has no definite colours. But even if such a thing were conceivable, it would again be something totally different from the intuitive idea of the lapse of time, to which the idealistic assertion refers. (p. 558, fn 5)

This is a pretty piece of ordinary language philosophizing. But like most of its ilk, it leaves one up in the air: even if one shares the linguistic intuitions, one can wonder how such intuitions can support weighty philosophical morals. To thrash through these issues, however, would lead us astray, for it is GTR and not STR that is true (or so we may suppose).

The relevance of GTR is it implies that the existence of matter causes the curvature of spacetime and thereby destroys the equivalence of inertial observers in the Minkowski spacetime presupposed in Gödel's argument from STR. Furthermore, in all the cosmological solutions to EFE known in the 1940s, there is a natural way to single out a distinguished time function. Gödel put it thus:

> The existence of matter, however, as well as the particular kind of curvature of space-time produced by it, largely destroy the equivalence of observers and distinguish some of them conspicuously from the rest, namely those which follow in their motion the mean motion of matter. Now in all cosmological solutions of the gravitational equations (i.e., in all possible universes) known at present the local times of all *these* observers fit together into one world time, so that apparently it becomes possible to consider this time as the "true" one, which lapses objectively. (p. 559)

From such considerations, Gödel noted, James Jeans had concluded, in Gödel's words, that "there is no reason to abandon the intuitive idea of an absolute time lapsing objectively" (p. 559).[28]

At this juncture Gödel issues a demurrer. In a footnote he mentions that the proposed method for picking out a preferred simultaneity can be challenged. A successful challenge would open the way to parroting within GTR the above argument from STR. Gödel notes that making the notion of the mean motion of matter into a precise concept may involve "introducing more or less arbitrary elements (such as, for example, the size of the regions or the weight function to be used in the computation of the mean motion of matter)." And he goes on to assert "It is doubtful whether there exists a precise definition which has so great merits, that there would be sufficient reason to consider exactly the time thus obtained as the true one (p. 560, fn 9)." One can dispute this claim; for example, for the class of models Jeans and Gödel had in mind, one may be able to prove that there exists a unique family of time slices with minimal intrinsic curvature. Such a family would arguably be a good candidate for defining the true time. I will not pursue this matter since Gödel does not seem to put much weight on his demurrer to Jeans and bases his case-in-chief on other considerations.

So the dialectical situation is now this. According to Gödel, STR supports the thesis of the ideality of time. However, Gödel acknowledges that proponents of an objective lapse of time can, with some justice, claim that GTR supports their case. To break the stalemate Gödel proposes to turn the tables by showing how an expanded knowledge of the solutions to EFE establishes his ideality thesis. His strategy is based on his discovery of a new class of solutions to EFE. In these solutions "the aforementioned procedure of defining an absolute time is not applicable, because the local time of the special observers used above cannot be fitted together into one world time" (p. 560). What Gödel meant is that there are solutions to EFE where matter is everywhere rotating so that the natural way of singling out a time function by taking the spacelike hypersurfaces orthogonal to the world lines of matter is not available. The technical point is this. If V^a is the normed timelike tangent field of the congruence of world lines of matter, the *rotation* or *twist* of the congruence is defined by $\omega_{ab} =: \nabla_{[a} V_{b]}$. By Frobenius' theorem, the congruence is hypersurface orthogonal just in case $\omega_{ab} = 0$. The point can be seen in a non-technical way by analogizing the world lines of matter to the strands of a rope. If the rope is twisted, the strands will not be orthogonal (in the Euclidean sense) to any plane slicing through the rope.[29] Granting for sake of argument that in a universe where matter is everywhere rotating there is no natural way to single out a distinguished time function, how does this conclusion bear on time in the actual universe where, presumably, matter on the average has no twist? I will not press the point here since a similar one will soon arise.

Gödel does not rest his case with his twist argument but goes on to claim that the idealistic viewpoint is strengthened by some of the other surprising

features of his solution. The considerations that flow from these features add up to something less than an argument. The task is to locate and assess the missing premise(s) that would produce a valid argument.

(P1) In the Gödel universe there is no global time function, nor does there exist a single global time slice, nor is there a globally consistent time order.
(P2) Therefore, in the Gödel universe there is no objective lapse of time.
(P3) The Gödel model satisfies EFE and other conditions, such as non-negativity of energy density, that one wants for a physically possible model.
(P4) Furthermore, the Gödel model cannot be excluded a priori on the grounds that time travel leads to the grandfather paradox.
(P5) However, the Gödel universe can be excluded a posteriori as a model for the actual universe since, for example, it gives no cosmological redshift.
(P6) But our universe is different from the Gödel universe only because of contingent features—in particular, the distribution and motion of matter.
(P7) ?
(C) Therefore, time in our universe is ideal.

Premises (P1), (P3), (P5), and (P6) are uncontroversial. And given Gödel's analysis of time lapse, (P2) is also unexceptionable. (P4), however, is controversial. In its defense Gödel writes:

> This and similar contradictions [i.e., the grandfather paradox], however, in order to prove the impossibility of the worlds under consideration, presuppose the actual feasibility of the journey into one's own past. But the velocities which would be necessary in order to complete the voyage in a reasonable length of time are far beyond everything that can be expected ever to become a practical possibility. (p. 561)

On the analysis of the grandfather paradox I have offered, it seems to me that Gödel's way of dismissing the grandfather paradox is too quick. On the other hand, my analysis does support the contention that cosmological models with CTCs and the other features listed in (P1) cannot be easily dismissed as conceptually or physically impossible worlds. Thus, the evaluation of Gödel's argument devolves to the question of what has to go into (P7) in order to make (C) follow from (P1) through (P7).

Here is one try.

(P7.1) The existence of an objective lapse is not a property that time can possess contingently.

This way of filling in (P7) is supported by the following passage from Gödel's essay:

> It might, however, be asked: Of what use is it if such conditions [i.e., those of (P1)] prevail in certain *possible* worlds? Does that mean anything for the question interesting us whether in *our* world there exists an objective lapse of time? I think it does. . . . The mere compatibility with the laws of nature of worlds in which there is no distinguished absolute time, and, therefore, no objective time lapse can exist, throws some light on the meaning of time also in those worlds in which absolute time *can* be defined. For, if someone asserts that this absolute time is lapsing, he accepts the consequence that, whether or not an objective lapse of time exists . . . depends on the particular way in which matter and its motion are arranged in the world. This is not a straightforward contradiction; nevertheless, a philosophical view leading to such consequences can hardly be considered as satisfactory. (pp. 561–562)

The most direct and the crudest interpretation of the pattern of argument would be: $L \rightarrow \mathcal{N}(L)$ (if time has the property of lapsing, then necessarily so), $\neg\mathcal{N}(L)$ (lapsing is not a necessary property of time), therefore $\neg L$ (time does not lapse). $\neg\mathcal{N}(L)$ is equivalent to $P(\neg L)$, where $P(\cdot)$ means that $\cdot$ is possible. And $P(\neg L)$ is established by showing there is a physically possible world—the Gödel universe—where $\neg L$ is true.

Gödel's essentialist intuitions here are not easy to fathom. There seems to be no lurking contradiction or anything philosophically unsatisfactory in saying in the same breath: "Space in the actual world is open, but if the mass density were a little greater, space would be closed," or "Time in the actual universe goes on forever into the future, but if the mass density were greater the universe would eventually recollapse and time would come to an end." Why then is there a lurking contradiction or something philosophically unsatisfactory in saying: "Time in our universe lapses, but if the distribution and motion of matter were different, there would be no consistent time order and so time would not lapse"? Gödel seemed to have thought that one should see the unsatisfactory character of this utterance just by reflecting on the concept of time. This game of using an inner sense to perceive conceptual truths is a dangerous one, for others claim to perceive the non-existence of CTCs as essential to the concept of time and, therefore, that contrary to (P4) the Gödel model can be ruled out on a priori grounds. Gödel gives us no guidelines for judging superiority of conceptual insight.

But there is an even more puzzling feature of Gödel's endorsement of $\neg L$. He concedes at this juncture of the dialectic that the actual universe has all the geometrical properties necessary for an objective time lapse, namely, the existence of an appropriately distinguished global time function. So in affirming $\neg L$ he must be claiming that time in the actual universe lacks some non-geometrical feature necessary for time lapse. What is this missing ingredient? Recall that Gödel says "The existence of an objective time . . . means (or, at least, is equivalent to the fact) that reality consists of an infinity of layers of 'now' which come into existence successively." By hypothesis the actual universe consists of an infinity of layers of 'now'. So Gödel must have believed that these 'nows' fail to "come into existence successively." There

are two mysteries here. First, how is the non-existence of a global time function in some other possible world relevant to whether the 'nows' of the distinguished time function of this world come into existence successively? Second, if what Gödel's argument for the ideality of time amounts to is that time lacks a shifting 'nowness', then there is no need to invoke GTR and the Gödel universe. For even if the actual universe and all physically possible universes were fully Newtonian, it would be difficult to make any non-psychologistic sense of shifting nowness.

Another attempt to fill in (P7) comes from Yourgrau (1991).

> (P7.2) "Since the actual world is lawlike compossible with the Gödel universe, it follows that our direct experience of time is ***compatible*** with its ideality. . . . But if even ***direct experience*** is inadequate to establish the existence of . . . genuine, successive time that lapses or passes—then ***nothing further*** will suffice." (Yourgrau 1991, p. 53)

In support of this reading Yourgrau cites the following passage from Gödel's essay:

> If the experience of the lapse of time can exist without an objective lapse of time, no reason can be given why an objective lapse of time should be assumed at all. (p. 561)

Let it be granted for sake of argument that some observers in the Gödel universe are under an illusion—they experience a time lapse and in consequence think that time objectively lapses even though in fact there is no objective lapse of time. How is that fact about the Gödel universe supposed to impinge on us? Granted, it should make us cautious in drawing consequences about the lapsing of time from our own experiences. ***But apart from our experiences of time lapse we have all sorts of other experiences that lend strong support to the inference that we do not inhabit a Gödel type universe but rather a universe that fulfills all of the geometrical conditions necessary for an objective lapse of time.***

To block this move, consider

> (P7.3) There are cosmological models that (i) lack the features necessary for an objective time lapse, but (ii) reproduce the redshift, etc., so that they are effectively observationally indistinguishable from models that fit current astronomical data and have the spatio-temporal structure needed to ground an objective lapse of time.

This tack is suggested by another passage from Gödel's essay:

> Our world, it is true, can hardly be represented by the particular kind of rotating solutions referred to above (because these solutions are static and,

> therefore, yield no red-shift for distant objects); there exist however also *expanding* rotating solutions. In such universes an absolute time might fail to exist, and it is not impossible that our world is a universe of this kind. (p. 562)

And indeed Gödel did go on to generalize his solutions to EFE in such a way as to allow for a cosmological redshift (Gödel 1952). However, I very much doubt there are cosmological models which allow for time travel and which are observationally indistinguishable from non-time travel models.[30] But even if there were, Gödel would seem to need an additional premise asserting something like a verifiability theory of meaning in order to reach his conclusion (C). I take it that few people will be attracted by such a premise.

I have been unable to locate any plausible argument which starts from Gödel's considerations and leads to the conclusion that time is ideal. Rather, what I find is a collection of arguments each of which is intriguing but ultimately unpersuasive. A bunch of unpersuasive arguments do not add up to one persuasive one. Reading between the lines of his "Reply to Criticisms," one can infer that this was Einstein's view also. Einstein begins by praising Gödel's essay as "an important contribution to the general theory of relativity, especially to the analysis of the concept of time" (Einstein 1949b, p. 687). But then he immediately brushes aside the question of the relation of GTR to idealistic philosophy and goes on to discuss issues of causation. This seems to me to be the correct response to Gödel.

Notes

1. Here is a sample: Boulware (1992); Carroll et al. (1992); Charlton and Clarke (1990); Cutler (1992); Deser (1993); Deser et al. (1992); Deutch (1991); Echeverria et al. (1991); Friedman and Morris (1991a, 1991b); Friedman et al. (1990); Friedman et al. (1992); Frolov (1991); Frolov and Novikov (1990); Gibbons and Hawking (1992); Goldwirth et al. (1993); Gott (1991); Grant (1993); Hawking (1992); Kim and Thorne (1991); Klinkhammer (1992); Menotti and Seminara (1993); Mikeeva and Novikov (1992); Morris et al. (1988); Novikov (1989, 1992); Ori (1991a, 1993); Ori and Soen (1994); Politzer (1992, 1994); 't Hooft (1992); Thorne (1991); Visser (1993, 1994); Yurtsever (1990, 1991).

2. A representative sample of this literature follows: Brown (1992); Chapman (1982); Dummett (1986); Dwyer (1975, 1977, 1978); Ehring (1987); Harrison (1971); Horwich (1989); Lewis (1986); MacBeath (1982); Mellor (1981); Smith (1986); Thom (1975); and Weir (1988).

3. Although it is a truism, it needs repeating that philosophy of science quickly becomes sterile when it loses contact with what is going on in science.

4. This appellation is suggested by some passages in H. G. Wells' *Time Machine* (1968), but I do not claim to have captured what Wells meant by time travel. For a review of and references to some of the science fiction literature on time travel, see Gardner (1988).

5. One way to make (C0) precise is to require that there exists on M a continuous, non-vanishing, timelike vector field.

6. Good treatments of Gödel's model are to be found in Malament (1985), Pfarr (1981), and Stein (1970). Gödel's original (1949a) cosmological model is a dust-filled universe, i.e., $T^{ab} = \mu U^a U^b$ where μ is the density of the dust and U^a is the normed four-velocity of the dust. This model is a solution to Einstein's field equations only for a non-vanishing cosmological constant $\Lambda = -\omega^2$, where ω is the magnitude of the rotation of matter. Alternatively, the model can be taken to be a solution to EFE with $\Lambda = 0$ and $T^{ab} = \mu U^a U^b + (\omega^2/8\pi) g^{ab}$; however, this energy–momentum tensor does not seem to have any plausible physical interpretation. Gödel's original model gives no cosmological redshift. Later Gödel (1952) generalized the model to allow for a redshift.

7. Recall from chapter 3 that a *time slice* is a spacelike hypersurface without edge.

8. Recall from chapter 3 that a *partial Cauchy surface* is a time slice that is achronal, i.e., is not intersected more than once by any timelike curve.

9. Some philosophers apparently think time travel is logically or conceptually impossible; see Hospers (1967, p. 177) and Swinburne (1968, p. 169).

10. Sometimes the local-to-global property may fail in causally nice spacetimes because singularities develop in solutions to some field equation. But one may regard such a failure as indicating that the field is not a fundamental one; see chapter 3.

11. The need for such a distinction has been previously noted by Bryson Brown (1992).

12. There would be an easy victory here if it were the case that in a world which is nomologically accessible from the actual world and which has a spacetime structure M, g_{ab} containing CTC, it is a law that the spacetime is M, g_{ab}. But on no account of laws with which I am familiar will it be the case that in, say, a Gödelian universe it is a law that the spacetime is Gödelian. However, it could be complained that in effect we are treating the spacetime structure as lawlike in taking it to be a fixed background on which to solve for consistency constraints; but this complaint seems to me to be another version of that in the first therapy (T1).

13. For a review of the horizon problem and attempted solutions, see chapter 5.

14. The slogan here is that there is no difference in laws without some difference in non-nomic facts. Some authors take this supervenience thesis to be a necessary truth (i.e., to hold for all possible worlds). David Lewis (1986, pp. ix–xii) takes it to be a contingent truth which holds only for possible worlds near the actual world.

15. However, the matter is complictaed by the fact that Huygens' principle (see chapter 5) does not hold in curved spacetimes.

16. Gödel took the possibility of time travel to support the conclusion that time is "ideal." Gödel's argument is discussed in an appendix to this chapter.

17. For more on this distinction, see Earman (1994).

18. *The weak energy condition* ($T_{ab} V^a V^b \geq 0$ for any timelike V^a) entails the *null energy condition* ($T_{ab} K^a K^b \geq 0$ for any null K^a). Taken together, the weak energy condition and EFE entail the *null convergence condition* ($R_{ab} K^a K^b \geq 0$ for any null K^a), which does the real work in Hawking's theorem.

19. Ori (1993) and Ori and Soen (1994) argue that a time machine solution can satisfy the weak energy condition right up to the time when the chronology violation starts.

20. For more details, see Earman (1994).

21. See Yurtsever (1991) for a definition of this notion and for a more precise specification of the stability property.

22. See Yurtsever (1991) for a formulation of the relevant smoothness conditions.

23. The space of initial conditions has a natural topology so that a generic set of solutions can be taken to be one that corresponds to an open set of initial conditions.

24. This condition requires that $\int_\gamma T_{ab}K^aK^b\,d\lambda > 0$ where $K^a = dx^a/d\lambda$ is the tangent to the null geodesic γ and λ is an affine parameter of γ.

25. The energy density depends on the behavior of the two-dimensional cross-sectional area of a pencil of geodesics.

26. Results of this character have been announced by Klinkhammer and Thorne (1990) in a never published preprint; see Echeverria et al. (1991).

27. Valuable information about the development of Gödel's ideas on time are to be found in Malament (1995) and Stein (1990, 1995).

28. The reference here is to Jeans (1936). Jeans claimed that in GTR "time regained a real objective existence, although only on an astronomical scale, and with reference to astronomical phenomena" (p. 22).

29. I owe this illustration to David Malament (1995). By examining the early versions of Gödel's manuscripts for (1949b), Malament was led to hypothesize that Gödel started his investigation by searching for solutions to EFE with rotating matter and subsequently discovered that his particular rotating solution contained CTCs.

30. At least if the CTCs cannot be unwound by passing to a covering spacetime, as is the case with the Gödel universe. For more on observational indistinguishability, see chapter 7.

31. My assessment of Gödel's argument for idealism was developed in correspondence with Steven Savitt. Our ideas are entangled but somewhat different. The reader is urged to consult Savitt (1994).

7

Eternal Recurrence, Cyclic Time, and All That

7.1 Introduction

The idea of a cyclic or repeating time finds an astonishingly broad acceptance in the history of thought, being found among such diverse and widely separated peoples as the Stoics of ancient Greece, the Hindus of India, the Taoists of China, and the Mayans of Central America.[1] Christianity was generally hostile to the idea since it clashed with the doctrine of the uniqueness of Christ.[2] The scientific revolution produced a decidedly ambivalent stance. On the one hand, Newtonian mechanics seemed to hold out the possibility of a "clockwork" universe. But on the other hand Newton himself worried that because of the instability of the solar system God would from time to time have to wind up the clock. Nineteenth and early twentieth century physics produced another tension. Poincaré's recurrence theorem showed that a closed mechanical system will almost surely return to approximately the same state; but the second law of thermodynamics seemed to indicate that the universe is destined to wind down to a heat death.

Doing full justice to this complex subject would require a separate book-length treatment. My aim is not so high; indeed, my remarks here will be narrowly confined to the implications of GTR and, more particularly, to the obstacles that spacetime singularities and the censorship of naked singularities pose to eternal recurrence. It is necessary at the outset to distinguish between two related but different ideas which are sometimes confused: first, the idea that numerically distinct but otherwise similar states occur over and over again in an open time; and second, the idea that the universe progresses through a series of changes only to return to the numerically identical state. Although terminology differs in these matters, I will use *eternal recurrence* to refer to the former, and *circular*, *cyclic*, or *closed time* to refer to the latter.

7.2 Tolman on eternal recurrence

The classical form of the second law of thermodynamics implies an increase in entropy with increasing time and thus appears to be in conflict with the idea that the universe can undergo regular periodic changes. However, Richard C. Tolman's investigation of the extension of thermodynamics to general relativity convinced him that the relativistic version of the second law permitted thermodynamic change to take place "at a finite rate entirely reversibly without any increase in entropy at all" (Tolman 1931b, p. 1759; see also 1931a). Tolman was thus led to consider the possibility of the periodic behavior of a non-stationary universe. He found, however, that for the simple case of a universe filled with a uniform distribution of matter, his requirements for thermodynamic reversibility clashed with the conditions for periodic behavior in the sense of continual expansions and contractions between fixed finite limits. From the modern perspective it is easy to see the problem even if thermodynamical considerations are set aside. As mentioned in chapter 5, the symmetry properties of a homogeneous and isotropic universe, which Tolman was assuming, force the stress–energy tensor to have the form of a perfect fluid: $T^{ab} = (\mu + p)U^aU^b + pg^{ab}$, where μ is the matter density and p is the pressure. If $\mu > 0$, $p \geq 0$, and the cosmological constant Λ is set to 0 (as Tolman assumed), then EFE entail that for the FRW universe with $k = +1$ (closed space sections with constant positive curvature), whose line element is

$$ds^2 = a^2(t)\left[\frac{dr^2}{1-r^2} + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)\right] - dt^2, \qquad (7.1)$$

the scale factor $a(t)$ ("radius of the universe") must go to 0 at some time in the past or future. The same conclusion holds if $\mu + 3p > 0$ and Λ is not too large.

Having excluded truly periodic behavior, Tolman turned to what he termed "quasi-periodic behavior" where the volume for a spatially finite universe increases from 0 to an upper bound and subsequently contracts to 0, and so on ad infinitum. The solutions he found "could not be regarded analytically as strictly periodic" (Tolman 1931b, p. 1765). Again the reason is easy to see from the standard treatment in terms of the FRW models. In these models the early stages of the universe are assumed to be radiation dominated. Then as a consequence of EFE (for $\Lambda = 0$), the scale factor $a(t)$ behaves approximately as $t^{1/2}$. So as the big bang ($t = 0$) is approached, $a(t) \to 0^+$, $\dot{a}(t) \to +\infty$, and $\ddot{a}(t) \to -\infty$, which are hardly the analytical conditions for a minimum for $a(t)$. A similar remark applies equally to the big crunch.

Nevertheless, Tolman continued to speak of quasi-periodic behavior.

> It is evident physically that contraction to zero volume could only be followed by another expansion, and in addition, as noted by Einstein in a similar connection, the idealization on which our considerations have been based can be regarded as failing in the neighborhood of zero volume. Hence from a physical point of view it seems reasonable to consider that solutions of the kind which we are considering could correspond to a series of successive expansions and contractions. (Tolman 1931b, p. 1765)

Similar sentiments are expressed in a joint article with Morgan Ward published in the following year (Tolman and Ward 1932, pp. 841–842). In both instances Tolman was wavering between two ideas. The first is that when the idealizations of perfect homogeneity and isotropy are removed, the singularities can be avoided and strictly periodic behavior can be achieved. For a time Einstein also harbored the hope that spacetime singularities would disappear when the idealizations were removed (see chapter 1). This hope was expressed, for example, in Einstein (1931), the paper to which Tolman referred in the above quotation. I will have more to say on this matter in section 7.4 below. The second idea is that even if the idealizations are accepted and the universe contracts to zero volume "it is evident physically that contraction to zero volume could only be followed by another expansion." But on reflection there is nothing physically evident in the notion that the contraction to zero volume would result in the sudden reversal of the sign of $a(t)$, followed by renewed expansion. Indeed, as I will now argue, the notion of the universe recycling itself though a succession of big bangs and big crunches is physically meaningless for the models under consideration.

7.3 Extending through the big bang and the big crunch

The most straightforward, if not the only sensible way, to cash in the notion of the universe recycling itself is to exploit the concept of the extension of a spacetime (as explained in chapter 2). Here we seem to quickly run into a dead end rather than a recycling. It was noted that for the radiation-dominated regime, which holds near $t = 0$, $a(t) \sim t^{1/2}$. As a result the FRW metric is not extendible as a C^2 (or even C^{2-}) metric. This follows from the fact that the Kretschmann curvature scalar $K(t) =: R_{abcd}(t)R^{abcd}(t)$ blows up rapidly as $t \to 0^+$.

But why should the extended metric be C^2? If continuity/ differentiability (c/d) conditions on the metric are relaxed, is it possible that the FRW metrics can be extended through the big bang? The answer to the second question is trivially yes if no c/d conditions are put on the extended metric. In answer to the first question, there may be no sufficient reason to require that the extended metric be C^2; but there should be enough c/d to assure that EFE make sense at least distributionally (see chapter 2). If not, then although a mathematically well-defined extension through the big bang or the big crunch may exist, it can justifiably be deemed physically meaningless. In chapter 2 it was argued that the requirement in question is implemented by the

condition that the metric be *regular* in the sense of Geroch and Traschen (1987). This condition is enough to prove some negative results.

The metric (7.1) lives on the manifold $M = S^3 \times (0, b)$, $(0, b) \subset \mathbb{R}$, $0 < b < +\infty$. Try to imagine that M is imbedded by the natural inclusion map into $\tilde{M} = S^3 \times (a, b)$, $-\infty < a < 0$. (This seems to have been Tolman's implicit assumption.) This means that the coordinates r, θ, ϕ, t are C^∞ for $a < t < b$. In these coordinates the contravariant components $g^{rr}, g^{\theta\theta}, g^{\phi\phi}$ of the metric are not locally bounded at $t = 0$ since they behave as $1/t$.[3] This does not settle the matter since we have to deal with the possibility that $M = S^3 \times (0, b)$ can be imbedded as a proper subset in a larger $\hat{M}$ in such a fashion that the r, θ, ϕ, t coordinates break down at the boundary $t = 0$. This would open up the possibility that in some new coordinate system $\hat{r}, \hat{\theta}, \hat{\phi}, \hat{t}$ which belongs to the C^∞ atlas of $\hat{M}$ but which is related to the r, θ, ϕ, t coordinates by a transformation which is singular at $t = 0$, the components of the FRW metric remain locally bounded at $t = 0$. It is hard to envision how this sort of behavior could occur. But the example of the Schwarzschild metric discussed in chapter 2 should give one pause; for it took decades before general relativists realized that an imbedding exists in which the Droste coordinates break down at the Schwarzschild radius in such a way that a smooth continuation of the Schwarzschild metric is possible. I conjecture that the FRW case is not analogous to the Schwarzschild case in this respect, but since I have no proof, I turn to a different and weaker demonstration of the inextendibility of the FRW metric.

Not only does one want EFE to make sense distributionally, one also wants the local conservation law $\nabla_a T^{ab} = 0$ to make sense in the same way. To obtain the conservation law as a consequence of EFE, the Bianchi identity $\nabla_{[a} R_{bc]d}{}^{e} = 0$ is needed. And for this identity to be meaningful in the sense of distributions it is necessary to require not only that the metric be regular but also that $R_{abc}{}^{d}$ be locally square integrable.[4] It follows that the absolute value of the Kretschmann curvature scalar K must be locally integrable. But if integrability is judged with respect to the FRW metric this is not the case if one tries to extend to $t = 0$. For example, in the $k = 0$ case, $|K(t)| \sim 1/a^8(t)$. In the radiation-dominated era shortly after the big bang, $a(t) \sim t^{1/2}$ so that $|K(t)| \sim 1/t^4$. Using the volume measure from the FRW metric, the volume integral of $|K(t)|$ is proportional to $t^{-5/2}$, which diverges as $t \to 0^+$. Of course, it is possible to define other volume measures with respect to which $R_{abc}{}^{d}$ is locally square integrable at the big bang; but unless some physical significance can be assigned to such measures, the physical significance of the extension remains moot.[5]

To the extent that this argument is successful, a similar argument can be used to show that it is not physically meaningful to extend through other scalar polynomial curvature singularities with strong blowup behavior. But as we know from chapter 2, there are many other types of spacetime singularities, and it is not at all evident whether and how the non-existence of physically meaningful extensions can be proved in the general case.

If the assumption of radiation dominance in the early universe is dropped in favor of matter dominance, then another route is opened to the conclusion that there is no physically meaningful extension through the big bang. For then the big bang singularity will be of the strong curvature type discussed in section 3.5; namely, for every causal geodesic approaching the big bang, all volume forms tend to 0. So in any extension that is C^0 every physical object would be crushed to zero volume. Arguably such an extension is not physically meaningful since the identity of physical objects is lost (see Ori 1991b).

7.4 Finding God in the big bang

It has been charged that attempts to find a prior cause of the big bang, whether physical or metaphysical, are incoherent. The gist of the objection is that the search for a prior cause assumes the existence of instants of time prior to the big bang, an assumption incompatible with the cosmological models that are the basis for our belief in the big bang. Thus, Adolf Grünbaum has written:

> To suggest or assume tacitly that instants existed after all before the big bang is simply *incompatible* with the physical correctness of the putative big bang model at issue, and thus implicitly denies its soundness. . . . It is now clear that the physical correctness of this model is also implicitly denied by someone who addresses any of the following questions to it: "What happened *before* the big bang?", "What prior events *caused* matter to come *into* existence at $t = 0$?", "What prior events caused the big bang to occur at $t = 0$?" . . . it is *altogether wrongheaded* . . . to complain that—even when taken to be physically adequate—the putative big bang model *fails to answer questions* based on assumptions which it denies as false. (Grünbaum 1991, pp. 238–239)

I must demur slightly from these opinions. By itself, a model of the big bang—say, a standard FRW model—is neither compatible nor incompatible with the notion that there are instants of time before the initial singularity. The fate of that notion depends on our choice of extendibility conditions. I have argued above that under plausible constraints on what is to be counted as a physically meaningful extension, there are no physically meaningful extensions through the big bang of the standard models. Perhaps the reader will find my argument convincing, perhaps not. But my first point is that some such argument is needed. My second point is that even if my argument, or some other, succeeds, it remains open that there is some mathematically meaningfully extension—involving lower continuity/differentiability conditions than those required for a physically meaningful extension—and that God or some other metaphysical cause operates in this mathematical time. I will return to this point below.

William Lane Craig (1991, 1994) has contended that even if it is conceded there is no meaningful sense in which there are moments of time before the big bang, the theist who wants to see God at work in the big bang still has options available (see also Craig and Smith 1993). Thus, it is claimed that the Creator's act of causing the universe to begin to exist could be conceived to be simultaneous with the universe's beginning to exist. Or He could be conceived to "exist timelessly and to cause tenselessly the origin of the universe at the Big Bang singularity" (Craig 1994, p. 329). But if the argument of section 7.3 is effective in showing that there is no physically meaningful extension to times before the big bang, it is equally effective in showing that there are no meaningful extensions to $t = 0$.

Why is it that theists want to find God in the big bang? Of course, if one is determined to find God, He can be found everywhere. But what in particular is it about the big bang, as opposed, say, to a flower, that makes His presence evident? Here is an excerpt of a recent letter to *The Wall Street Journal* which encapsulates one answer to our question that I have found to be not uncommon in popular thought:

> Reductionist science has made vast strides in the past 400 years, but it has now hit the brick wall of the Big Bang. Most physicists admit their 'standard model' cannot explain it. One example: If all the mass of the universe were packed into that one dimensionless point just before the Big Bang, gravitational attraction would be intense beyond comprehension. How then to explain a momentary suspension of that attractive force, suddenly reversed into an incredibly huge explosive force, followed at once by the re-emergence of gravity? An act of God perhaps?[6]

Rather than locating an instant of time at which God does his work, what the author of this letter actually provides is a sketch of a reductio of the assumption that there is a spacetime event corresponding to a state where all the mass of the universe is "packed into that one dimensionless point." The technical argument of the preceding section confirms the reductio.

A seemingly more sophisticated but not essentially different response is that something cannot begin to exist without a cause; so if there is no physical cause of the beginning to exist, there must be a metaphysical one. Here I am in complete agreement with Professor Grünbaum in that the standard big bang models are not compatible in any obvious way with the idea that the universe has a physically uncaused beginning. Indeed, these models imply that for every time t there is a prior time t' and that the state at t' is a cause (in the sense of causal determinism) of the state at t.[7] Craig (1994) has responded that a beginning for time does not require that time have a first instant but only that time be finite in the past.[8] So for Craig the FRW models of the big bang exhibit a temporal structure in which time began to exist even though there is no first instant of time. Consequently, for Craig the principle *Whatever begins to exist has a cause* applies to these models. However, on Craig's reading this principle is not an obvious "metaphysical truth"; in particular, it is not a consequence of the widely held principle *Every event has a cause*, which is satisfied in the FRW big bang models without any help from the theists.

As explained in chapter 2, there are ways of doctoring standard relativistic cosmological models to represent singularities as boundary points attached to the spacetime manifold. In this way one could hope to reclaim a "first instant" for time, opening the way for an application of the principle *Every event has a cause* and the invocation of a deistic cause in the absence of a physical cause. But in the most widely cited method of doctoring—the b-boundary approach—the big bang is represented as a single point that is not Hausdorff-separated from interior points.[9] This counterintuitive consequence serves to emphasize the fact that the boundary points are ideal elements, a warning that remains in effect even if some other means of adding the boundary points that escapes the counterintuitive features of the b-boundary construction were to be found. Nothing prevents the theist from seeing God as operating at these ideal points. But since ideal points are not points of spacetime, the sense in which God can be said to cause or bring about the universe by operating at these points is very remote from the usual causal notions of science and everyday life that are concerned with connections between events in space and time. This is not to say that theistic talk about God creating the universe is illegitimate. But it is to say that such talk finds no special purchase in the big bang. *Even in models with no big bang and with time extending infinitely far into the past, ideal points corresponding to $t = -\infty$ could be attached to the spacetime manifold and God's helping hand could be seen at work there.*

I have a parallel reaction to the complaint that the scientific models of the big bang leave much unexplained—why, for example, the universe began with the matter content it did and not some other, and why the expansion from the big bang obeys EFE and not some other empirical equations (see Quinn 1993). The point behind the complaint is perfectly correct: science leaves unexplained the most fundamental laws it has been able to uncover, and it cannot say why one rather than another of the myriad histories compatible with these laws has been actualized. But again, this observation applies to all scientific modeling of natural phenomena, not just to the big bang models. As far as I can see, the big bang models offer no special advantages to the deists.

Speaking purely personally now, it strikes me as bordering on the sacrilegious to see God's creative force as able to operate only at a singularity or ideal point. It is more to His glory if He operates everywhere and everywhen, and if He operates independently of such contingencies as whether there is an initial singularity and, if so, what type it is. Those who want to find God in the big bang should beware of falling into the trap of relegating God to the diminishing interstices left by modern science. Once the trap is recognized it is easy to escape using God's supernatural attributes. If there is no first instant for the physical universe or no prior physical time to the big bang at which God can operate, no matter. The Creator "may be conceived

to exist in a metaphysical time" and thus "to exist temporally prior to the inception of physical time" (Craig 1994, p. 328). The constraints of physics cannot bind the Creator. But precisely to the extent that a supernatural cause of the beginning of the universe does not have to answer to the constraints of nature, scientists qua scientists are entitled to ignore it.

7.5 No recurrence theorems

After this digression into theology, let us return to the question of eternal recurrence. In section 7.3 it was argued, contra Tolman, that eternal recurrence cannot be achieved in the context of standard big bang–big crunch models. But as Tolman noted, these models involve very unrealistic idealizations. It remains to be seen whether no recurrence results can be demonstrated under more realistic conditions.

We must first investigate the possibility of cheap counterexamples to no recurrence. EFE admit a wide class of *stationary* solutions. Technically, stationarity of the metric g_{ab} means that there is a timelike vector field **V** such that $\pounds_{\mathbf{V}} g_{ab} = 0$, or equivalently, $\nabla_{(a} V_{b)} = 0$.[10] (Such a **V** is called a *Killing field*.) This is the invariant way of saying that the spacetime metric does not change with time since it guarantees that for any spacetime point there is a neighborhood covered by a coordinate system $\{x^1, x^2, x^3, t\}$ such that $V^a = (\partial/\partial t)^a$ and $\partial g_{ab}/\partial t = 0$. But by itself stationarity is not enough for a counterexample to no recurrence. Gödel spacetime (see chapter 6) is stationary, but there is no reasonable sense in which this spacetime exhibits eternal recurrence; indeed, Gödel spacetime does not contain any global time slices so that it does not make sense either to affirm or deny that in this setting things are the same at two different "instants."

To remedy this defect we can require the spacetime to be *static*, that is the spacetime is not only stationary but also the Killing field **V** is hypersurface orthogonal, i.e., $V_{[a} \nabla_b V_{c]} = 0$.[11] But staticity is still not sufficient to produce a counterexample since it is compatible with closed and cyclical time structure, which will be studied in section 7.6. So let us also posit that the spacetime possesses an open time structure by requiring it to admit a time function t whose level surfaces coincide with the orthogonal hypersurfaces of the timelike Killing field and whose value increases along every future-directed timelike curve. Such a spacetime is trivially periodic, periodic for every period, at least with respect to the spacetime geometry since the geometry of any orthogonal hypersurface of **V** is the same as any other. (Following along the trajectories of **V** gives the isometry of one hypersurface onto another.)

Even after all this fiddling, two more things are needed to produce a counterexample to no recurrence. First, it is necessary not only that the spatial geometry but also that all the physical fields on spacetime be the same at the two instants in question. Of course, in any cosmological model M, g_{ab}, T^{ab} that satisfies EFE, the energy–momentum tensor T^{ab} will inherit the symmetries of the metric so that in a stationary spacetime where $\pounds_{\mathbf{V}} g_{ab} = 0$, it will also be the case that $\pounds_{\mathbf{V}} T^{ab} = 0$. But it is not automatic that the fields that generate T^{ab} inherit the symmetries of the metric. For example, when T^{ab} is generated by an electromagnetic field, there are stationary solutions to the Einstein–Maxwell equations for which the Maxwell tensor F^{ab} characterizing the electromagnetic field is not stationary, i.e., $\pounds_{\mathbf{V}} F^{ab} \neq 0$. It has also been conjectured that there are static solutions to the Einstein–Maxwell equations where the electromagnetic field does not inherit the symmetry, but no specific example is known (see Michalski and Wainwright 1975). Thus, in searching among the static solutions of EFE for a counterexample to no recurrence one needs either to confine attention to vacuum solutions ($T^{ab} = 0$)[12] or else one must check that the source fields are in fact static.

Second, to assure that the static solution is a genuine counterexample to no recurrence, the solution should be maximal. This rules out the possibility that the solution is merely a piece of a larger spacetime that is non-static. That the worry here is a real one is illustrated by the exterior Schwarzschild solution, which is a static vacuum solution. But this solution is extendible, and its maximal analytic extension—Kruskal spacetime—is non-stationary.

Even with all of these caveats in place, counterexamples to no recurrence can be found among the static solutions to EFE. The static Einstein universe (see Hawking and Ellis 1973, p. 139) is a case in point. However, this cosmological model relies on a non-zero value for the cosmological constant Λ. And it is also unstable; for a fixed value of Λ, the smallest change in the mass density will cause the universe to evolve. From here on I propose to concentrate on solutions to EFE with vanishing Λ. Then in searching for static counterexamples to no recurrence there are various candidates among spatially open models. Negative mass Schwarzschild spacetime is a spatially open, static, inextendible, vacuum solution to EFE with $\Lambda = 0$ (see chapter 3). But as the name indicates, it has the unphysical property of a negative ADM mass. It also violates cosmic censorship since the central singularity is naked. The truncated Schwarzschild spacetime of Janis, Newman, and Winicour (1968) is a static solution corresponding to a massless scalar field. It coincides with the exterior Schwarzschild solution for $r > \alpha$ but the Schwarzschild sphere $r = \alpha$ becomes a point singularity. This singularity is also naked. There are also static plane wave solutions, but these lack Cauchy surfaces (see Penrose 1965) and therefore violate strong cosmic censorship. Weyl (1917) produced a class of axisymmetric static solutions. But in those cases where maximal extensions have been obtained—e.g., the Curzon monopole solution (see Scott and Szekeres 1986a, b)—the resulting spacetime is either non-static or contains naked singularities. These examples raise the following question: Are there inextendible static solutions to EFE with $\Lambda = 0$ that do not violate cosmic censorship and are stable? Apparently the answer is not known. Recently it has been found that for a stress–energy source corresponding to a Yang–Mills field, there are solutions to EFE with $\Lambda = 0$

singularity free, asymptotically Minkowskian, and topologically Wasserman, Yau, and McLeod 1991; Smoller and these particle-like solutions the Yang–Mills repulsive gravitational attraction to achieve a static situation. are unstable.

a physically satisfactory static counterexample to no example would be of a very special sort and would ture that in a universe which harbors some change change never brings about a return to an exactly cannot prove or refute this conjecture, I turn to cases proving a general no recurrence theorem.

counterexamples to no recurrence seem to be spatially a tight no recurrence result may be found for spatially concretely, we can begin by abstracting some of the closed FRW models, which we know from section 7.2 eternal recurrence and, indeed, with any recurrence.[13] contain Cauchy surfaces. That is a feature we want to interesting form of recurrence the recurrence should evolution rather than by chance. For such cases recurrence coincide, since Laplacian determinism similar state recurs once, it will recur over and over know that the existence of a Cauchy surface implies manifold M is diffeomorphically $\Sigma \times \mathbb{R}$. And to realize recurrence we want to restrict attention to the case which expresses in a rigorous way the notion that the closed. Next, the standard energy conditions (see chapter $\Lambda = 0$) allow us to infer the convergence condition that non-spacelike V^a. And finally we may assume that all geodesics feel a tidal force at some moment so that the genericity conditions hold. All of this (plus some differenconditions on the metric) allows us to apply the Hawking–Penrose (see chapter 2) to conclude that the spacetime is timelike incomplete. The question then becomes whether the by the geodesic incompleteness prevent recurrence, as case.

reductio assume that recurrence happens. Then as already implies that recurrence happens infinitum. Tipler (1980) recurrence plus the compactness of the space sections of a certain complete timelike geodesic. Then by the constructions used in the classic Hawking–Penrose (see chapter 2), he showed that the existence of such a contradiction. This is an important result but it does not answer to our question. The constructions of the theorems require that the metric be C^2.[14] But as argued conditions for a physically meaningful spacetime may

be lower than C^2, perhaps low enough so as not to preclude that in some non-C^2 but physically meaningful extension eternal recurrence takes place. Of course, it may be that the singularities demonstrated in Tipler's result are strong enough that they cannot be removed by a physically meaningful extension, as I argued was the case for the FRW big bang models. But until this is shown, we do not have a really solid argument against eternal recurrence for generic, spatially closed, deterministic models.[15]

Under some further conditions, which imply the uniqueness and stability of the Cauchy development of the initial data for EFE, Tipler (1980) was also able to establish that for the class of spatially finite models in question, not only is return to an exactly similar state impossible but so is return to a state arbitrarily close to the initial one.[16] This result shocks intuitions trained on classical mechanics, where the famous Poincaré recurrence theorem shows that a spatially bounded system with a finite number of degrees of freedom and finite energy almost always returns arbitrarily closely to the initial state.[17] It is clear that the GTR is much less hospitable than is classical physics to the notion of eternal return and that the reasons have to do with the formation of spacetime singularities. But because of uncertainties about how to characterize essential singularities and also because of the difficulties in proving the existence of such singularities in a general setting, it is unlikely that we will have in hand in the near future a precise knowledge of just how inhospitable GTR is to eternal recurrence.

7.6 Cyclic time

As noted in section 7.1, eternal recurrence should not be confused with a cyclic time structure.[18] As used here the former refers to the recurrence of similar states in an open time, while the latter refers to return to numerically the same state in a circular or closed time. It will become apparent, however, that there is an intimate connection between eternal recurrence and cyclic time. But first it is important to be more precise about what it means for time to be open or, alternatively, to be closed.

One approach would be to relativize the open versus closed distinction to observers. Thus, if we idealize an observer as an inextendible future-directed timelike curve γ, we could say that for observer $O(\gamma)$ time is open (respectively, closed) just in case γ is open (respectively, closed). For garden variety spacetime this definition entails that for every observer time is open. In Gödel spacetime the definition entails that time for any non-accelerated observer is open, whereas time for some sufficiently strongly accelerated observers is closed. Although it is hard to fault this definition on its own terms, it does seem overly narcissistic. For instance, in the cylindrical spacetime of Fig. 7.1, time for $O(\gamma)$ is open since γ never returns to the same spacetime location. But there is also a clear intuitive sense in which time itself in this universe should be counted as circular.

which are static, singularity free, asymptotically Minkowskian, and topologically $\mathbb{R}^4$ (see Smoller, Wasserman, Yau, and McLeod 1991; Smoller and Wasserman 1993). In these particle-like solutions the Yang–Mills repulsive force balances the gravitational attraction to achieve a static situation. However, these solutions are unstable.

Even if there were a physically satisfactory static counterexample to no recurrence, the counterexample would be of a very special sort and would not refute the conjecture that in a universe which harbors some change (non-stationarity), the change never brings about a return to an exactly similar state. Since I cannot prove or refute this conjecture, I turn to cases where there is hope of proving a general no recurrence theorem.

All of the potential counterexamples to no recurrence seem to be spatially open. This suggests that a tight no recurrence result may be found for spatially closed universes. More concretely, we can begin by abstracting some of the features of the spatially closed FRW models, which we know from section 7.2 are incompatible with eternal recurrence and, indeed, with any recurrence.[13] The FRW models contain Cauchy surfaces. That is a feature we want to preserve since for an interesting form of recurrence the recurrence should result from deterministic evolution rather than by chance. For such cases recurrence and eternal recurrence coincide, since Laplacian determinism entails that if an exactly similar state recurs once, it will recur over and over again ad infinitum. We know that the existence of a Cauchy surface implies that the spacetime manifold M is diffeomorphically $\Sigma \times \mathbb{R}$. And to realize our hope of proving no recurrence we want to restrict attention to the case where Σ is compact, which expresses in a rigorous way the notion that the universe is spatially closed. Next, the standard energy conditions (see chapter 3) and EFE (with $\Lambda = 0$) allow us to infer the convergence condition that $R_{ab}V^aV^b \geq 0$ for any non-spacelike V^a. And finally we may assume that all timelike and null geodesics feel a tidal force at some moment so that the timelike and null genericity conditions hold. All of this (plus some differentiability conditions on the metric) allows us to apply the Hawking–Penrose singularity theorem (see chapter 2) to conclude that the spacetime is timelike or null geodesically incomplete. The question then becomes whether the singularities indicated by the geodesic incompleteness prevent recurrence, as they do in the FRW case.

For purposes of reductio assume that recurrence happens. Then as already noted, determinism implies that recurrence happens infinitum. Tipler (1980) showed that this eternal recurrence plus the compactness of the space sections allows the construction of a certain complete timelike geodesic. Then by the same conjugate point constructions used in the classic Hawking–Penrose singularity theorems (see chapter 2), he showed that the existence of such a geodesic leads to a contradiction. This is an important result but it does not give an unassailable answer to our question. The constructions of the Hawking–Penrose theorems require that the metric be C^2.[14] But as argued in chapter 2, the c/d conditions for a physically meaningful spacetime may

be lower than C^2, perhaps low enough so as not to preclude that in some non-C^2 but physically meaningful extension eternal recurrence takes place. Of course, it may be that the singularities demonstrated in Tipler's result are strong enough that they cannot be removed by a physically meaningful extension, as I argued was the case for the FRW big bang models. But until this is shown, we do not have a really solid argument against eternal recurrence for generic, spatially closed, deterministic models.[15]

Under some further conditions, which imply the uniqueness and stability of the Cauchy development of the initial data for EFE, Tipler (1980) was also able to establish that for the class of spatially finite models in question, not only is return to an exactly similar state impossible but so is return to a state arbitrarily close to the initial one.[16] This result shocks intuitions trained on classical mechanics, where the famous Poincaré recurrence theorem shows that a spatially bounded system with a finite number of degrees of freedom and finite energy almost always returns arbitrarily closely to the initial state.[17] It is clear that the GTR is much less hospitable than is classical physics to the notion of eternal return and that the reasons have to do with the formation of spacetime singularities. But because of uncertainties about how to characterize essential singularities and also because of the difficulties in proving the existence of such singularities in a general setting, it is unlikely that we will have in hand in the near future a precise knowledge of just how inhospitable GTR is to eternal recurrence.

7.6 Cyclic time

As noted in section 7.1, eternal recurrence should not be confused with a cyclic time structure.[18] As used here the former refers to the recurrence of similar states in an open time, while the latter refers to return to numerically the same state in a circular or closed time. It will become apparent, however, that there is an intimate connection between eternal recurrence and cyclic time. But first it is important to be more precise about what it means for time to be open or, alternatively, to be closed.

One approach would be to relativize the open versus closed distinction to observers. Thus, if we idealize an observer as an inextendible future-directed timelike curve γ, we could say that for observer $O(\gamma)$ time is open (respectively, closed) just in case γ is open (respectively, closed). For garden variety spacetime this definition entails that for every observer time is open. In Gödel spacetime the definition entails that time for any non-accelerated observer is open, whereas time for some sufficiently strongly accelerated observers is closed. Although it is hard to fault this definition on its own terms, it does seem overly narcissistic. For instance, in the cylindrical spacetime of Fig. 7.1, time for $O(\gamma)$ is open since γ never returns to the same spacetime location. But there is also a clear intuitive sense in which time itself in this universe should be counted as circular.

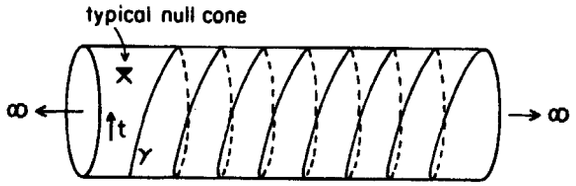

Fig. 7.1 An observer with an open world line in a temporally closed spacetime

To get away from the narcissism of particular observers we could look at a congruence $\mathscr{C}$ of inextendible timelike curves (i.e., through every point of the spacetime there passes exactly one member of $\mathscr{C}$). Since we are dealing with time orientable spacetimes, such a congruence always exists. We could then say that relative to $\mathscr{C}$, time in M, g_{ab} is open (respectively, closed) just in case every $\gamma \in \mathscr{C}$ is open (respectively, closed). The Gödel universe is dust filled. The world lines of the dust particles are timelike geodesics and, therefore, are open. Thus, the world lines of the dust form a congruence $\mathscr{C}$ relative to which time in the Gödel universe is (according to the present definition) open. There is no alternative $\mathscr{C}'$ relative to which time in the Gödel universe is (according to the present definition) closed. In the cylindrical spacetime of Fig. 7.1 there is obviously a congruence of closed timelike curves. But only slightly less obviously, there is also a congruence of open timelike curves. The latter is open to the charge of group narcissism. It is true that no world line of this group returns to the same event in spacetime, but however one partitions the cylinder by connected time slices, each world line of the group in question returns to the same slice—which argues that time can only plausibly be considered as closed in this example. A related objection to the present approach is that, as will be illustrated below, there are spacetimes which can plausibly be considered to have a closed temporal structure even if there are no CTCs.

The defender of the observer-oriented approach can shrug off these objections. But for purposes of capturing the idea of a cyclic time structure in the sense of return to the same state, a definition couched in terms of time slices seems inescapable. The following idea then suggests itself.

DEFINITION 7.1
A time-oriented spacetime M, g_{ab} has an *open time* structure just in case there is a *linear time function*, i.e., a map $t: M \to \mathbb{R}$ where (i) t is C^0, (ii) for every $r \in \mathbb{R}$ in the range of t, $t^{-1}(r)$ is a time slice, and (iii) for any $p, q \in M$ such that $p \ll q$, $t(p) < t(q)$.[19]

One way to justify this definition is to verify that the *quotient topology* is in fact $\mathbb{R}$. To explain what this means, suppose that $\equiv$ is an equivalence relation on the topological space X. The quotient topology assigned to the equivalence classes $X/\equiv$ is the finest topology in which the projection map $\pi: X \to X/\equiv$

is continuous. In the present case, the equivalence relation for the linear time function t is defined by the condition that for $p, q \in M$, $p \equiv q$ just in case $t(p) = t(q)$. To show that the associated quotient topology is indeed $\mathbb{R}$ it suffices to show that t is an open map. Towards this end, note that if Def. 7.1 holds, sets of the form $I^-(q) \cap I^+(p)$ form a basis for the manifold topology of M (see Hawking and Ellis 1973, pp. 196–197). It is easy to verify that $t(I^-(q) \cap I^+(p)) = (t(p), t(q))$ and, consequently, that t is an open map.

Instead of Def. 7.1 we could have started with

DEFINITION 7.2
A time-oriented spacetime M, g_{ab} has an *open time structure* just in case there is a collection $\mathscr{S}$ of time slices that partition M and the quotient topology of $M/\mathscr{S}$ is $\mathbb{R}$.

We have already seen that Def. 7.1 entails Def. 7.2. The converse is not hard to establish.

In parallel with Defs. 7.1 and 7.2 one can try to characterize a closed or circular time structure in the following two ways.

DEFINITION 7.3
A time-oriented spacetime M, g_{ab} has a *closed time structure* just in case there is a *circular time function*, i.e., a map $u: M \to S^1$ where (i) u is C^0, (ii) for any $s \in S^1$, $u^{-1}(s)$ is a time slice, and (iii) for any distinct $p, q, y, z \in M$, if there is a future-directed timelike curve which goes from p to q to y to z and which does not reintersect the slice $u^{-1}(u(p))$, then $u(p), u(y)$ separate $u(q), u(z)$ on the circle S^1.

DEFINITION 7.4
A time-oriented spacetime M, g_{ab} has a *closed time structure* just in case there is a collection $\mathscr{S}$ of time slices that partition M and the quotient topology $M/\mathscr{S}$ is S^1.

Just as Defs. 7.1 and 7.2 are equivalent, so are Defs. 7.3 and 7.4.

One potential drawback to the present approach is that it is not applicable to many acausal spacetimes—Gödel spacetime, for example, which does not admit any time slices. However, the conclusion that Gödel spacetime cannot be taken to have either an open or a closed time structure strikes me as correct. Similarly, it seems to me correct to conclude in line with Defs. 7.1 through 7.4 that the spacetime of Fig. 7.2, which can be partitioned by time slices, is temporally neither open nor closed.

Somewhat more surprising and harder to swallow is the fact that according to Def. 7.1 through 7.4 there are spacetimes that can be considered to be both temporally open and closed. A relevant example is illustrated in Fig. 7.3 where two half-lines have been removed from the two-dimensional cylindrical spacetime of Fig. 7.1. If one likes, by introducing an appropriate

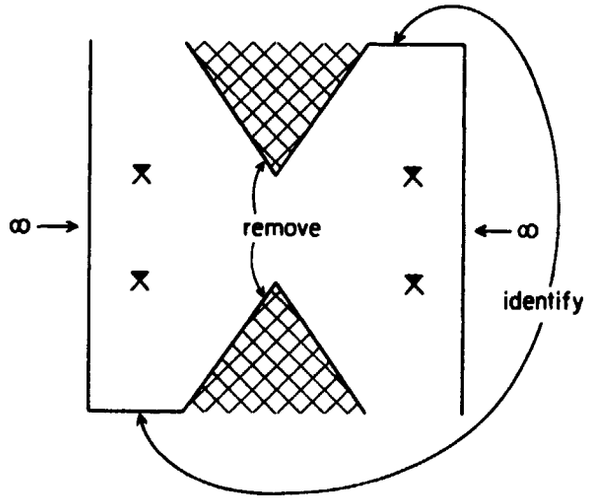

Fig. 7.2 A spacetime that is partitioned by time slices but is temporally neither open nor closed

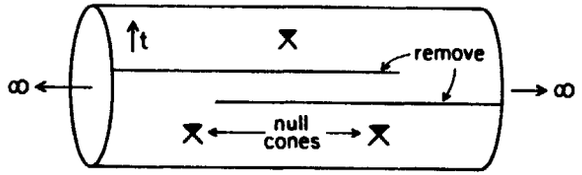

Fig. 7.3 A spacetime that may be considered to be both temporally open and closed

conformal factor in the metric, the lines can be taken to represent irremovable curvature singularities. The removed lines overlap sufficiently that the spacetime is *stably causal* (i.e., a slight widening of the light cones does not result in CTCs; see chapter 6). By a theorem of Hawking and Ellis (1973, Prop. 6.4.9) the spacetime admits of a global time function so that Def. 7.1 is fulfilled and time can be considered open.[20] On the other hand there is a more obvious partition $\mathscr{S}$ by time slices such that the quotient topology $M/\mathscr{S}$ is S^1.

An even nastier surprise comes from Robert Geroch's[21] observation that Defs. 7.3 and 7.4 allow us to count Minkowski spacetime as having a circular time structure. To see this, choose the partition $\mathscr{S}$ such that the level surfaces of some inertial time coordinate that lie at constant multiples of some chosen unit are counted as components of the same slice. One peculiar feature of this slicing is that the slices are not achronal. But achronality is not in general a reasonable condition to impose in cases of circular time structures; for instance, achronality is violated in the spacetime of Fig. 7.1, which certainly counts as a paradigm case of a circular time structure. Another notable feature of the Geroch slicing of Minkowski spacetime is that the slices have disconnected components; in fact, each slice has a countably infinite number of them. Similarly, some of the level surfaces of any linear time function for the spacetime of Fig. 7.3 will be disconnected. Thus, requiring connectedness of the time slices would with a single stroke get rid of both of the counterintuitive examples. But the price to be paid for soothing our intuitions is too high. For example, in the case of two-dimensional Minkowski spacetime with a closed ball removed, it seems natural to count the level surfaces of an inertial time coordinate as good time slices, even though some of them will be disconnected. If alternatively the disconnected components are regarded as different time slices, then in this spacetime the topology of time, as characterized by the quotient topology, will have a branching and then recoalescing structure which is non-Hausdorff. The former alternative, which counts the topology of time as $\mathbb{R}$, seems preferable.

My proposal is that we keep Defs. 7.1 and 7.2 as they are and that we simply swallow the consequence that time in the spacetime of Fig. 7.3 can be considered open. By contrast, however, I find it intolerable for time in Minkowski spacetime to be counted as closed. Thus, while Defs. 7.3 and 7.4 supply necessary conditions for a closed time, they cannot be considered to give sufficient conditions. The clue to an appropriate strengthening lies in the fact that those examples where we have no compunction about saying that a spacetime exhibits a closed time structure result from making identifications in a covering spacetime with an open and periodic time structure. If we make this feature the definition of a closed time structure, then Defs. 7.3 and 7.4 will, of course, be satisfied, but Minkowski spacetime will no longer be counted as exhibiting a closed time structure. I thus propose to replace Defs. 7.3 and 7.4 with

DEFINITION 7.5
A time-oriented spacetime M, g_{ab} has a *closed time structure* just in case M, g_{ab} results from identifying the time slices modulo $\Delta \in \mathbb{R}$ in a temporally open and Δ-periodic covering spacetime.

If this proposal is accepted, we have the intimate connection, hinted at above, between eternal recurrence and cyclic time. But the connection serves to draw much of the interest of the latter; for a cyclic time structure is never intrinsic but only arises because the Great Topologist has made identifications in a larger spacetime. Furthermore, the various no-go results for eternal recurrence discussed in previous sections also tell against the physical possibility of cyclic time.

When he wrote *The Philosophy of Space and Time*, Hans Reichenbach thought that we always have the option of interpreting away a CTC in favor of an open timelike curve on which similar events are repeated over and over ad infinitum (see Reichenbach 1958, pp. 140–143, 272–273). This belief is mistaken if the interpreting away takes the form of unwinding the CTC in a covering spacetime. As we know from chapter 6, Gödel spacetime is a relevant

example; it contains CTCs, but because it is simply connected it is its own universal covering spacetime. But given the present proposal for defining a closed time structure, Reichenbach's position has more merit for cases where CTCs inhabit a spacetime with a closed time structure. But it still remains to ask whether the two descriptions—a cyclic time structure in the base spacetime versus an open but periodic time structure in a covering spacetime—are equivalent in a sense so strong that the choice between them can be regarded as conventional.

At a minimum we would want it to be the case that the two spacetimes are observationally indistinguishable. Clark Glymour (1977) has proposed a criterion for observational indistinguishability which runs as follows:

DEFINITION 7.6
The spacetimes M, g_{ab} and M', g'_{ab} are *observationally indistinguishable* just in case for every inextendible timelike curve γ of M, g_{ab} there is an inextendible timelike curve γ' of M', g'_{ab} such that $I^-(\gamma)$ and $I^-(\gamma')$ are isometric, and similarly with M, g_{ab} and M', g'_{ab} interchanged.

This definition uses the idealization, already employed above, of observers as inextendible timelike curves,[22] and it assumes that 'observation' is to be given a causal reading so the events that an observer can in principle observe are precisely those within her past light cone.[23] Then the idea is that no observer will be able to tell which of the two spacetimes she inhabits just in case for any observer of the one spacetime there is an observer of the other such that the past light cones of the two are the same in the sense of being isometric.[24] Applying Def. 7.6 to examples of spacetimes with closed time structures yields mixed results for Reichenbach. The spacetime of Fig. 7.3 possesses a closed time structure according to my proposed analysis in Def. 7.5; and according to Def. 7.6, it is observationally indistinguishable from its universal covering spacetime. But this is an untypical case since the spacetime of Fig. 7.3 also possesses an open time structure. More typical are cases like that of Fig. 7.1 where the spacetime and its universal covering are not observationally distinguishable according to Def. 7.6. On behalf of Reichenbach it can be argued that although the spacetimes in question are in principle observationally distinguishable, no observer can be in a position to know which of the two she inhabits, at least not if mind–body identity is true; for by construction the corresponding states of the two spacetimes are identical as regards the value of any physical quantity[25] and, thus, the corresponding mental states that supervene on the body states of the observers must also be the same.

7.7 Conclusion

For those interested in exotic time structures, GTR proves to be something of a tease. To illustrate how time can "branch," one can cite examples of general relativistic spacetimes that can be partitioned by a collection of time slices where initially the slices are connected but later become disconnected. By regarding the disconnected components as different time slices, the quotient topology becomes a branching "Y". But typically the disconnected pieces can be uniformly regarded as different components of the same slices so that the quotient topology is the standard $\mathbb{R}$; indeed, if Penrose's form of the cosmic censorship hypothesis (see chapter 3) holds, then there is a partition by time slices in which all of the slices are connected.

Mathematical examples of general relativistic spacetimes illustrating eternal recurrence and cyclic time structures can also be given. But because of the occurrence of singularities the physics of GTR is distinctly unfriendly towards the former; and if the proposed characterization of cyclic time structure as a rolling up of open periodic spacetime is accepted, GTR is equally unfriendly towards the latter. Just how unfriendly GTR can be is a matter that cannot be resolved until we settle some delicate questions about the nature of essential singularities.

One interesting by-product of the analysis is that the open time versus closed time dichotomy is not a dichotomy, for there are spacetimes that exhibit neither structure and others that exhibit both. In this sense the global structures of general relativistic spacetimes are more exotic than popular and philosophical imaginations have been able to grasp. According to the analysis given here, a spacetime can be considered to have a closed time structure even though it does not contain CTCs. But in those cases where CTCs are involved, all the problems about causality discussed in chapter 6 are implicated.

Notes

1. For a nice overview and references to the literature, see Tipler (1980).

2. In "The City of God" (Bk XIII, Ch. 13) St. Augustine responded to the idea of "cycles of time," in which there should be a "constant renewal and repetition of the order of nature" as follows: ". . . far be it, I say, from us to believe this. For since Christ died for our sins; and, rising from the dead, He dieth no more" (Augustine 1948, pp. 191–192).

3. Nor is there any other admissible coordinate system in which all the contravariant components are locally bounded. For if the components g^{ij} are locally bounded in the coordinate system $\{x^i\}$ in the atlas of C^∞ charts for M, then likewise the components g'^{ij} in any other coordinate system $\{x'^i\}$ belonging to the atlas are locally bounded. (This is just to say that the requirement that g_{ab} and g^{ab} be locally bounded is an invariant one.) The transformation rule for the contravariant components of the metric is $g'^{ij} = (\partial x'^i/\partial x^m)(\partial x'^j/\partial x^n) g^{mn}$. The $\partial x'^k/\partial x^p$ are certainly locally bounded if the x'^k are C^∞ functions of the x^p. So if the g^{mn} are locally bounded, g'^{ij} is the sum of products of locally bounded functions and, thus, is itself locally bounded.

4. At least if the distribution must arise from a locally integrable tensor field; see chapter 2.

5. But nothing prevents God from using such a measure to peer through the big bang. The following section indulges in a bit of theology.

6. Letters to the Editor section, *Wall Street Journal*, June 16, 1993.

7. This point was made forcefully by Torretti (1979).

8. What are the necessary and sufficient conditions for a spacetime M, g_{ab} to have a temporal structure that is finite in the past? The reader may want to think about why the following is not a sufficient condition: every past-directed timelike half-curve has finite proper length.

9. Perhaps theists will find this feature attractive since it means that although God operates at the beginning of time, He is nevertheless near every event.

10. $\pounds_{\mathbf{V}}$ denotes the *Lie derivative* with respect to $\mathbf{V}$. $\pounds_{\mathbf{V}} g_{ab} = V^c \nabla_c g_{ab} + g_{cb} \nabla_a V^c + g_{ac} \nabla_b V^c$. This is equal to $\nabla_a V_b + \nabla_b V_a$ since $\nabla_c g_{ab} = 0$.

11. Equivalently, the *rotation* or *twist tensor* $\omega_{ab}(\mathbf{V})$ of $\mathbf{V}$ vanishes; see chapter 6.

12. It is typically assumed that the vanishing of T^{ab} implies the vanishing of the source fields. I will follow this assumption here.

13. In an FRW universe that expands from a zero volume and then recontracts to a zero volume, there will be a time $t = t^*$ of maximal expansion. The time slices $t = t^* \pm d$, $d > 0$, situated symmetrically about $t = t^*$ carry the same intrinsic spatial geometry. But this is not a counterexample to no recurrence since the extrinsic curvatures of the slices are different. If ξ^a is the normed timelike vector field orthogonal to the $t =$ constant slices, the *extrinsic curvature* of the slices is defined by $K_{ab} =: \nabla_a \xi_b$. At $t = t^* - d$ the universe is expanding ($K^a{}_a > 0$) while at $t = t^* + d$ the universe is contracting ($K^a{}_a < 0$). In a static spacetime the extrinsic curvature of the orthogonal hypersurfaces of the timelike Killing field vanishes.

14. Actually, as noted in chapter 2, C^{2-} will suffice.

15. I differ here with Torretti (1983, p. 337, note 26) on the substance and significance of Tipler's no recurrence theorem.

16. The same caveats expressed above about extendibility apply here as well. The relevant notion of closeness of states is made precise by using Sobolev spaces; see Tipler (1980).

17. For a system of point masses, the closeness of states is measured in the natural Euclidean metric on the classical phase space.

18. This section relies on and at the same time corrects some of my ideas in Earman (1977).

19. Recall that $p \ll q$ means that there is a future-directed timelike curve from p to q. The present definition of a linear time function differs only slightly from the definition of a global time function. For a global time function t it is required that t be differentiable and that $\nabla^a t$ be a timelike vector field. It follows that t is strictly increasing along timelike curves (clause (iii) of Def. 7.1) and that the level surfaces of t are spacelike hypersurfaces without edges (clause (ii) of Def. 7.1). Def. 7.1 is adopted here in order to have a symmetry with the definition of a circular time function given below.

20. The reader may wish to try to draw the level surfaces of a global time function for the spacetime of Fig. 7.3.

21. Private communication.

22. The reader may want to consider the implications of modifying Def. 7.6 so as to reflect the fact that actual observers only live for a finite amount of time; see Malament (1977).

23. For some caveats about this dogma, see chapter 5.

24. One might want to strengthen Def. 7.6 by requiring not only that $I^-(\gamma)$ and $I^-(\gamma')$ are isometric (which guarantees that the two are geometrically the same) but also that all physical fields are the same at the corresponding points of the isometry. As noted above, EFE do not guarantee that the latter follows from the former.

25. At least if Def. 7.6 is strengthened in the manner indicated in note 24.

8

Afterword

In a paper addressed to the measurement problem in QM, J. S. Bell and M. Nauenberg wrote:

> It seems that the quantum mechanical description will be superseded. In this it is like all theories made by man. But to an unusual extent its ultimate fate is apparent in its internal structure. It carries in itself the seeds of its own destruction. (Bell and Nauenberg 1966, p. 285)

Do spacetime singularities signal that classical GTR contains the seeds of its own destruction? Recall from chapter 1 Peter Bergmann's report of his and Einstein's opinions:

> It seems that Einstein always was of the opinion that singularities in classical field theory are intolerable. They are intolerable because a singular region represents a breakdown of the postulated laws of nature. I think one can turn this argument around and say that a theory that involves singularities and involves them unavoidably, moreover, carries within itself the seeds of its own destruction. (Bergmann 1980, p. 156)

Couple these opinions with the Hawking–Penrose singularity theorems, and the seeds of self-destruction in GTR bloom.

It seems to me, however, that as regards self-destruction there is an important difference between QM and GTR. The measurement problem in QM shows, prima facie, that the theory is empirically inadequate in the worst way: it cannot account for the fact that measurement procedures yield definite outcomes. As hope fades that some clever interpretational ploy will resolve this problem, it becomes more likely that some overhaul of quantum dynamics will be needed and, thus, that Bell and Nauenberg's moral is correct. By contrast, there is no correspondingly blatant failure of classical GTR to save the phenomena in connection with its prediction of singularities in the gravitational collapse of stars and in the big bang origin of the universe. In the former case, GTR's predictions about black holes appear to be gaining empirical successes. In the latter case, there are puzzles connected with the particle horizons of the standard big bang model. But as argued in chapter 5, these puzzles are to a large extent concerned not with empirical adequacy but with the nature of scientific

explanation; and in any case, the most popular fixes for the horizon problem do not banish the initial singularity.

Of course, the acceptability of a scientific theory concerns much more than the ability of the theory to save the phenomena. And it is evident that for many physicists singularities present a formidable obstacle to the acceptance of classical GTR. The first issue of this year's *General Relativity and Gravitation* contains an article entitled "No More Spacetime Singularities?"; it pronounces that "even as a classical theory, general relativity is deficient as a theory of spacetime because it predicts the existence of singularities" (Kostelecký and Perry 1994, p. 7).[1] But what exactly is it about singularities that makes GTR "deficient"? The authors do not say. In the Misner–Thorne–Wheeler bible, *Gravitation*, Charles Misner speaks of the "abhorrence" of the theoretical prediction of infinite curvature and infinite density which is "particularly heightened by the correlative prediction that these infinities occurred at a finite proper time in the past, and would—if they recur—occur again at some finite proper time in the future" (p. 813).[2] Relativity theory, in both its special and general forms, implies all sorts of things that seem abhorrent to intuitions trained on Newtonian physics. But generally the conclusion to be drawn is not that something is wrong with relativity theory but rather that intuitions need retraining. What then is it about spacetime singularities that calls for a modification of the theory rather than intuitions?

As if responding to this question, John Wheeler opined in *Gravitation* that the singularities of gravitational collapse confront physics with "its greatest crisis ever" (p. 1198). What is this crisis? For Wheeler it is encapsulated in the "paradox" of gravitational collapse: GTR says "'This is the end' and physics says 'there is no end'" (p. 1197), or "'collapse ends physics'; 'collapse cannot end physics'" (p. 1198). Granted, there is a superficial paradox in saying "The laws of GTR entail the existence of singularities," and then adding in the same breath "And in doing so they entail their own demise." But the air of paradox here is due to a loose way of speaking. For the singularities entailed by the theory do not exist at any spacetime location, and there is no event in spacetime where the laws fail to hold (chapter 2).

Perhaps this dissolution of Wheeler's paradox is too facile, for it is achieved by refusing to count as part of spacetime any location where the metric and, hence, the laws of GTR are not well defined. Perhaps then the residuum of a paradox lies in the fact that in entailing singularities GTR demonstrates its own incompleteness. This is the opinion of Brandenberger et al.

> Singularities are undesirable for a theory which claims to be complete, since their existence implies that spacetime cannot be continued past them. The spacetime structure becomes unpredictable already at the classical level . . . The presence of singularities is an indication that G[T]R is an incomplete theory. Wheeler even talks about a "crisis in physics." (Brandenberger et al. 1993, p. 1629)

In response to this "crisis" Brandenberger et al. propose to modify Einstein's field equations in such a way that singularities are avoided without sacrificing well-tested predictions of standard GTR.[3] But by its own lights GTR is not guilty of incompleteness because it entails black hole and big bang singularities. For either the singularities are essential or not. If they are inessential (i.e., removable), what's the beef? If they are essential and cannot be removed by any extension in which the laws of GTR make sense, then by the lights of the theory there is nothing further to be said. The theory is hardly convicted out of its own mouth of incompleteness for failing to answer questions about, for example, what happens 'before' the big bang and 'after' the big crunch, at least not if the theory implies, as argued in chapter 7, that such questions are not physically meaningful.

Perhaps the seeds of self-destruction are rooted not in incompleteness but in falsity. Perhaps, that is, in entailing singular behavior GTR is committed to empirically false predictions. But to sustain an interesting form of the seeds of its own destruction, the reason for thinking that GTR is committed to false predictions would have to be stated in the rubric of classical GTR. I know of no such reasons. Indeed, leaving aside possible quantum effects, GTR continues to pass every empirical test with flying colors (see Will 1993). A much different charge is that considerations from quantum physics suggest GTR breaks down in the vicinity of curvature singularities. Thus Robert Wald writes "The prediction of singularities undoubtedly represents a breakdown of general relativity in that its classical description of gravitation and matter cannot be expected to remain valid at the extreme conditions expected near a spacetime singularity" (Wald 1984a, p. 212). Behind Wald's remark is the idea that when curvature becomes sufficiently strong, quantum effects "which invalidate classical general relativity will play a dominant role" (ibid., p. 212). Of course, we know the singularities of GTR need not involve unbounded curvature (chapter 2). And even when they do, it remains a pious hope that some quantum theory of gravity, yet to be formulated, will contain mechanisms for the avoidance of singularities. Leaving aside such demurrers, the main point I wish to emphasize is that we have strayed far from the original contention that even taken on its own terms, classical GTR contains the seeds of its own destruction because it entails singularities. At least we have come far enough to break the analogy with the measurement problem in QM.

Having found no obvious merit in the charge that, even taken on its own terms, classical GTR stands convicted out of its own mouth of some heinous crime for pronouncing the existence of singularities, it is well to consider the opposite attitude that singularities are seeds of confirmation rather than seeds of self-destruction. After all, spacetime singularities are a feature that separates GTR from all of its predecessor Newtonian and special relativistic theories and from some of its competitor theories of gravitation.[4] So by confirming the existence of singularities, the theory would receive a big boost in empirical support. In addition, it is not hard to develop a fondness for certain types of

singularities. For example, a black hole singularity can be appreciated both as the ultimate garbage dump, able to take care of any waste disposal problem without the need to recycle, and as a source of extractable energy (see Wald 1984a, pp. 324–330).

There are two obstacles to this embrace of singularities. The first concerns how knowledge of the existence of singularities is to be achieved. To say that spacetime singularities exist is not to say that there are such and such events in spacetime whose presence can be detected, if only indirectly. (And the fact that we lack an attractive procedure for attaching singular points to the spacetime manifold speaks against talking about spacetime singularities as if they were objects, even ideal objects.) Rather to say that spacetime singularities exist—or better, that spacetime is singular—is to say that the large-scale structure of spacetime has such and such features, where the such and such features may be complicated and abstruse (chapter 2). But though difficult, the matter is not altogether desperate. By way of analogy, it is not easy to establish reasonably secure knowledge claims about other large-scale features of spacetime, for example, as to whether space is open or closed; for such claims must rely on a number of auxiliary theoretical assumptions, each of which stands in need of its own justification. And aside from a pervasive skepticism and antirealism, which would deny all knowledge claims outside of the realm of the directly observable, I do not see any reason in principle why claims about large-scale features of spacetime—including its singularity structure—cannot be established, at least by the standards that scientists use for evaluating other theoretical claims.

The second and more serious obstacle to the embrace of singularities is that no one wants to hug a naked singularity. (Exceptions: those with a taste for weirdness and those who want to perform supertasks; see chapter 4.) Indeed, if cosmic censorship fails for GTR, then it would seem that classical GTR is convicted out of its own mouth of the sin of incompleteness. At least the conviction is sustained *if* determinism is required for completeness for non-quantized theories, for violations of cosmic censorship are inextricably bound up with a breakdown in determinism (chapter 3). Perhaps one can also argue that violations of cosmic censorship would show that classical GTR is incomplete in a stronger sense. The premise required is not that determinism holds but the weaker premise that all physical processes be law governed. The argument would be completed by showing that classical GTR places no constraints, not even statistical ones, on what can emerge from a naked singularity. To date the evidence pro and con on whether GTR contains built-in mechanisms for enforcing cosmic censorship has been both scanty and mixed. Progress has been slow and probably will continue to be so owing to the difficulty in formulating and proving censorship theorems and in constructing counterexamples. It is to be hoped that the growing activity in numerical relativity will give us more insight into this crucial issue.

The presence of CTCs in solutions to Einstein's field equations can serve to buttress the charge that classical GTR is an incomplete theory insofar as

CTCs and other acausal features involve violations of cosmic censorship. Independently of cosmic censorship, CTCs would also support a separate charge of incompleteness *if* they were deemed to be conceptually or physically impossible, for then some selection principle over and above the laws of classical GTR would be needed to exclude them. However, it was argued in chapter 6 that the grandfather paradox and its ilk do not speak in favor of such an impossibility but are simply crude devices for bringing out the existence of consistency constraints entailed by the presence of CTCs. In some instances these constraints may aspire to law status. When the aspirations are fulfilled, GTR is not thereby shown to be incomplete for the actual world; rather, what is shown is that in some possible worlds which are nomologically accessible from the actual world and which contain CTCs, there are laws over and above those of classical GTR. But on an empiricist conception of laws this is hardly surprising since in traveling to other possible worlds—whether or not those worlds contain CTCs—one should be prepared to find that the laws are not the same as those of the actual world.[5]

Suppose for sake of discussion the reader is willing to seriously entertain my position that the fact GTR entails the existence of spacetime singularities need not mean it contains the seeds of its own destruction and that a generalized *horror singulariti* is not justified. How does it affect the search for a quantum theory of gravity? Obviously, the banishment of all spacetime singularities is no longer to be taken as a desideratum for quantum gravity. But beyond that obvious consequence the way forward is not clear. Although there may be no sound basis for a general horror of singularities, some types of singularities are accompanied by features that are justly cause for concern—the acausality of CTCs and the gross failure of determinism associated with naked singularities being the principle ones. For those for whom these concerns amount to alarm, quantum gravity looms as a savior. In chapter 6 it was seen that at present the prospects for proving chronology protection theorems in classical GTR seem dim.[6] Quantum gravity promises help since semiclassical calculations indicate that in some situations quantum fields diverge strongly on chronology horizons. Thus, one may hope that a full quantum theory of gravity will contain the mechanisms to prevent the manufacture of CTCs.[7] And one can also hope that if classical GTR lacks the resources to prevent the development of other types of naked singularities not associated with CTCs, then quantum gravity can supply the resources for censorship; at present, however, there is no clear basis for this latter hope.

Philosophers of science can be justly proud of their contributions to the foundations of QM and, in particular, to clarifying the measurement problem and to elucidating the meaning of the Bell inequalities. But thus far their meager efforts towards understanding the foundations of the other great theory of modern physics, GTR, and, in particular, towards understanding the problem of spacetime singularities does not merit any corresponding pride. This book is an initial effort to set out some of the many facets of the problem, to explain its intrinsic interest, and to indicate some of its implications for the

foundations of physics and the philosophy of science. It was written in the faith that, if adequately revealed, the problem of spacetime singularities will not remain the orphan of the philosophy of science and that if adopted as a rightful child it will enrich not only the philosophy of space and time but other members of the family as well. My faith can only be vindicated by the work of other, more able hands.

Notes

1. The authors go on to consider string theory as a way of avoiding singularities.

2. Misner goes on to consider ameliorating the problem by pushing the initial singularity into the infinite past; see Misner, Thorne, and Wheeler (1973, pp. 813–814).

3. A different proposal for achieving the same aim is given in Cornish and Moffat (1994).

4. Singularities are not very effective in separating GTR from other classical relativistic theories of gravity since in proving the existence of singularities Einstein's field equations are used in a weak way—essentially to derive the consequence that $R_{ab}V^aV^b \geq 0$ for any non-spacelike V^a; see chapter 2.

5. By an empiricist conception of laws I mean one which makes the laws of a world supervene on the occurrent features of that world; see chapter 6.

6. On this matter, see also Earman (1994).

7. Even if quantum gravity presents the manufacture of CTCs, it does not follow that quantum gravity is incompatible with CTCs in general. Quantum field theory on curved spacetimes provides some hints on this matter. Interacting quantum fields lose the property of unitarity in the presence of CTCs. However, a reasonable probability interpretation may still be possible (see Friedman et al. 1992).

References

Abbott, P. (ed) (1986). *Inflationary Cosmology*. Philadelphia: World Scientific.

Akdeniz, K. G., Arik, M., Mutus, A., and Rizaðglu, E. (1991). "Coasting Kaluza–Klein Cosmology," *Modern Physics Letters A*, 6, 1543–1546.

Albert, D. (1992). *Quantum Mechanics and Experience*. Cambridge, MA: Harvard University Press.

Albrecht, A., and Steinhardt, P. J. (1982). "Cosmology for Grand Unified Theories with Radiatively Induced Symmetry Breaking," *Physical Review Letters*, 48, 1220–1223.

Allis, V., and Koetsier, T. (1991). "On Some Paradoxes of the Infinite," *British Journal for the Philosophy of Science*, 42, 187–194.

Anderson, P. (1983). "Effects of Quantum Fields on Singularities and Particle Horizons in the Early Universe," *Physical Review D*, 28, 271–285.

———. (1984). "Effects of Quantum Fields on Singularities and Particle Horizons in the Early Universe. II," *Physical Review D*, 29, 615–627.

Arntzenius, F. (1992). "The Common Cause Principle," in D. Hull, M. Forbes, and K. Okruhlik (eds), *PSA 1992*, Volume 2. East Lansing, MI: Philosophy of Science Association, 227–237.

Augustine, St. (1948). "City of God," in W. J. Oates (ed), *The Basic Writings of Saint Augustine*, Volume II. New York: Random House, 3–663.

Barrow, J. D. (1988). "The Inflationary Universe: Modern Developments," *Quarterly Journal of the Royal Astronomical Society*, 29, 101–117.

———, and Silk, J. (1980). "The Structure of the Early Universe," *Scientific American*, 242 (April), 118–128.

———, and Tipler, F. J. (1986). *The Anthropic Cosmological Principle*. Oxford: Clarendon Press.

Bell, J. S., and Nauenberg, M. (1966). "The Moral Aspect of Quantum Mechanics," in A. De Shalit (ed), *Preludes in Theoretical Physics*. Amsterdam: North-Holland, 279–286.

Benacerraf, P. (1962). "Tasks, Super-Tasks, and the Modern Eleatics," *Journal of Philosophy*, LIX, 765–784.

———, and Putnam, H. (eds) (1983). *Philosophy of Mathematics*. 2nd ed. Cambridge: Cambridge University Press.

Bergmann, P. G. (1942). *Introduction to the Theory of Relativity*. New York: Prentice-Hall.

———. (1980). Remarks made included in "Open Discussion, Following Papers by S. W. Hawking and W. G. Unruh," in H. Woolf (ed), *Some Strangeness in the Proportion*. Reading, MA: Addison-Wesley, 156.

Birkhoff, G. D. (1923). *Relativity and Modern Physics*. Cambridge, MA: Harvard University Press.

Blau, S. K., and Guth, A. H. (1987). "Inflationary Cosmology," in S. W. Hawking and W. Israel (eds), *Three Hundred Years of Gravitation*. Cambridge: Cambridge University Press, 524–603.

Börner, G. (1988). *The Early Universe*. Berlin: Springer-Verlag.

Boulware, D. G. (1992). "Quantum Field Theory in Spaces with Closed Timelike Curves," *Physical Review D*, 46, 4421–4441.

Brandenberger, R., Mukhanov, V., and Sornborger, A. (1993). "Cosmological Theory without singularities," *Physical Review D*, 48 1629–1642.

Brans, C. H., and Randall, D. (1993). "Exotic Differentiable Structures and General Relativity," *General Relativity and Gravitation*, 25, 205–221.

Brown, B. (1992). "Defending Backwards Causation," *Canadian Journal of Philosophy*, 22, 429–444.

Budic, R., Isenberg, J., Lindblom, L., and Yasskin, P. B. (1978). "On the Determination of Cauchy Surfaces from Intrinsic Properties," *Communications in Mathematical Physics*, 61, 87–95.

Carroll, J. W. (1994). *Laws of Nature*. New York: Cambridge University Press.

Carroll, S. M., Farhi, E., and Guth, A. H. (1992). "An Obstacle to Building a Time Machine," *Physical Review Letters*, 68, 263–266.

Carter, B. (1971). "Causal Structure in Spacetime," *General Relativity and Gravitation*, 1, 349–391.

Chandrasekhar, S., and Hartle, J. B. (1982). "On Crossing the Cauchy Horizon of a Reissner–Nordström Black-Hole," *Proceedings of the Royal Society (London) A*, 348, 301–315.

Chapman, T. (1982). "Time Travel," and "Two Times," in *Time: A Philosophical Analysis*. Dordrecht: D. Reidel, 134–145.

Charlton, N., and Clarke, C. J. S. (1990). "On the Outcome of Kerr-Like Collapse," *Classical and Quantum Gravity*, 7, 743–749.

Chihara, C. S. (1965). "On the Possibility of Completing an Infinite Process," *Philosophical Review*, 74, 74–87.

Choquet-Bruhat, Y., and Geroch, R. (1969). "Global Aspects of the Cauchy Problem in General Relativity," *Communications in Mathematical Physics*, 14, 329–335.

———, De Witt-Morette, C., and Dillard-Bleich, M. (1982). *Analysis, Manifolds and Physics*. Revised edition. Amsterdam: North-Holland.

Chruściel, P. T. (1992). "On Uniqueness in the Large of Solutions of Einstein's Equations ('Strong Cosmic Censorship')," *Contemporary Mathematics*, 132. Providence: American Mathematical Society, 235–273.

———, and Isenberg, J. (1993). "Non-Isometric Vacuum Extensions of Vacuum Maximal Globally Hyperbolic Spacetimes," *Physical Review D*, 48, 1616–1628.

———, Isenberg, J., and Moncrief, V. (1990). "Strong Cosmic Censorship in Polarized Gowdy Spacetimes," *Classical and Quantum Gravity*, 7, 1671–1680.

Clarke, C. J. S. (1973). "Local Extensions in Singular Space-Times," *Communications in Mathematical Physics*, 32, 205–214.

———. (1975a). "The Classification of Singularities," *General Relativity and Gravitation*, 6, 35–40.

———. (1975b). "Singularities in Globally Hyperbolic Space-Time," *Communications in Mathematical Physics*, 41, 65–78.

———. (1977). "Time in General Relativity," in J. Earman, C. Glymour, and J.

Stachel (eds), *Foundations of Space-Time Theories, Minnesota Studies in the Philosophy of Science*, Volume VIII. Minneapolis: University of Minnesota Press, 94–108.

———. (1988). "Singularities, Problems and Prospects," in B. R. Iyer, A. Kembhavi, J. V. Narlikar, and C. V. Vishveshwara (eds), *Highlights in Gravitation and Cosmology*. Cambridge: Cambridge University Press, 15–29.

———. (1993). *Analysis of Space-Time Singularities*. Cambridge: Cambridge University Press.

———, and Krolak, A. (1985). "Conditions for the Occurrence of Strong Curvature Singularities," *Journal of Geometry and Physics*, 2, 127–143.

———, and Schmidt, B. G. (1977). "Singularities: The State of the Art," *General Relativity and Gravitation*, 8, 129–137.

Clauser, J. F., and Horne, M. A. (1974). "Experimental Consequences of Objective Local Theories," *Physical Review D*, 10, 526–535.

Collins, C. B., and Hawking, S. W. (1973). "Why is the Universe Isotropic?" *Astrophysical Journal*, 180, 317–334.

———, and Stewart, J. M. (1971). "Qualitative Cosmology," *Monthly Notices of the Royal Astronomical Society*, 153, 419–434.

Cornish, N. J., and Mofat, J. W. (1994). "Nonsingular Gravity Without Black Holes," *Journal of Mathematical Physics*, 35, 6628–6643.

Craig, W. L. (1991). "Pseudo-dilemma?," *Nature*, 254, 347.

———. (1994). "Professor Grünbaum on Creation," *Erkenntnis*, 40, 325–341.

———, and Smith, Q. (1993). *Theism, Atheism, and Big Bang Cosmology*. Oxford: Clarendon Press.

Crosswell, K. (1993). "The Quest for the Cosmological Constant," *New Scientist*, 137 (20 February), 23–27.

Curzon, H. E. J. (1924a). "Bipolar Solutions of Einstein's Gravitation Equations," *Proceedings of the London Mathematical Society*, Series 2, 23, xxix.

———. (1924b). "Cylindrical Solutions of Einstein's Gravitation Equations," *Proceedings of the London Mathematical Society*, Series 2, 23, 477–480.

Cutler, C. (1992). "Global Structure of Gott's Two-String Spacetime," *Physical Review D*, 45, 487–494.

De, U. K. (1969). "Paths in Universes Having Closed Time-Like Lines," *Journal of Physics A* (Series 2), 2, 427–432.

de Sitter, W. (1917a). "On the Curvature of Space," *Koninklijke Akademie van Wetenschappen te Amsterdam. Proceedings of the Section of Sciences*, 20, 229–242.

———. (1917b). "On Einstein's Theory of Gravitation and its Astronomical Consequences. Third Paper," *Monthly Notices of the Royal Astronomical Society*, 78, 3–28.

———. (1918). "Further Remarks on the Solutions of the Field-Equations of Einstein's Theory of Gravitation," *Koninlijke Akademie van Wetenschappen te Amsterdam. Proceedings of the Section of Sciences*, 20, 1309–1312.

Deser, S. (1993). "Physical Obstacles to Time-Travel," *Classical and Quantum Gravity*, 10, S67–S73.

———, Jakiw, R., and 't Hooft, G. (1992). "Physical Cosmic Strings Do Not Generate Closed Timelike Curves," *Physical Review Letters*, 68, 267–269.

Deutch, D. (1991). "Quantum Mechanics near Closed Timelike Lines," *Physical Review D*, 44, 3197–3217.

Dieckmann, J. (1988). "Cauchy Surfaces in Globally Hyperbolic Space-Times," *Journal of Mathematical Physics*, 29, 578–579.

Dominici, D., Holman, R., and Kim, C. W. (1983). "Horizon Problem in Brans-Dicke Cosmology," *Physical Review D*, 28, 2983–2986.

Droste, J. (1917). "The Field of a Single Center in Einstein's Theory of Gravitation, and the Motion of a Particle in That Field," *Koninklijke Akademie van Wetenschappen te Amsterdam. Proceedings of the Section of Sciences*, 19, 197–215.

Dummett, M. (1986). "Causal Loops," in R. Flood and M. Lockwood (eds), *The Nature of Time*. New York: Basil Blackwell, 135–169.

Dwyer, L. (1975). "Time Travel and Changing the Past," *Philosophical Studies*, 27, 341–350.

———. (1977). "How to Affect, but Not Change the Past," *Southern Journal of Philosophy*, 15, 383–385.

———. (1978). "Time Travel and Some Alleged Logical Asymmetries between Past and Future," *Canadian Journal of Philosophy*, 8, 15–38.

Eardley, D. M. (1987). "Naked Singularities in Spherical Gravitational Collapse," in B. Carter and J. B. Hartle (eds), *Gravitation in Astrophysics, Cargèse 1987*. New York: Plenum Press, 229–235.

Earman, J. (1977). "How to Talk About the Topology of Time," *Noûs*, 11, 211–226.

———. (1986). *A Primer on Determinism*. Dordrecht: D. Reidel.

———. (1989). *World Enough and Spacetime: Absolute vs. Relational Theories of Space and Time*. Cambridge, MA: MIT Press.

———. (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge, MA: MIT Press.

———. (1993). "In Defense of Laws: Reflections on Bas van Fraassen's *Laws and Symmetries*," *Philosophy and Phenomenological Research*, LIII, 413–419.

———. (1994). "Outlawing Time Machines: Chronology Protection Theorems," *Erkenntnis*. In press.

———, and Glymour, C. (1980). "The Gravitational Red Shift as a Test of General Relativity: History and Analysis," *Studies in History and Philosophy of Science*, 11, 175–214.

———, and Janssen, M. (1993). "Einstein's Explanation of the Motion of Mercury's Perihelion," in J. Earman, M. Janssen, and J. D. Norton (eds), *The Attraction of Gravitation: New Studies in the History of General Relativity, Einstein Studies*, Volume 5. Boston: Birkhäuser, 129–172.

———, and Norton, J. D. (1994). "Infinite Pains: The Trouble with Supertasks," to appear in S. Stich (ed), *Paul Benacerraf: The Philosopher and His Critics*. New York: Blackwell.

Echeverria, F., Klinkhammer, G., and Thorne, K. S. (1991). "Billiard Balls in Wormhole Spacetimes with Closed Timelike Curves; Classical Theory," *Physical Review D*, 44, 1077–1099.

Eddington, A. S. (1923). *The Mathematical Theory of Relativity*. Cambridge: Cambridge University Press.

———. (1924). "A Comparison of Whitehead's and Einstein's Formulae," *Nature*, 113, 192.

Ehring, D. (1987). "Personal Identity and Time Travel," *Philosophical Studies*, 52, 427–433.

Einstein, A. (1914). "Die Formale Grundlage der Allegemeiner Relativitätstheorie," *Königlich Preussische Akademie (Berlin). Sitzungsberichte*, 1030–1085.

———. (1915). "Erklärung der Perihelbewegung des Merkur aus der Allgemeinen Relativitätstheorie," *Königlich Preussische Akademie (Berlin). Sitzungsberichte*, 844–847.

———. (1916). "Die Grundlagen der Allgemeinen Relativitätstheorie," *Annalen der Physik*, 49, 769–822. Reprinted in English translation in W. Perrett and G. B. Jeffrey (eds), *The Principle of Relativity*. New York: Dover, 1957.

———. (1918). "Kritisches zu Einer von Hrn. De Sitter Gegebenen Lösung der Gravitationsgleichungen," *Königlich Preussische Akademie der Wissenschaften (Berlin). Sitzungsberichte*, 270–272.

———. (1927). "Allgemeine Relativitätstheorie und Bewegungsgesetz," *Preussische Akademie der Wissenschaften (Berlin). Physikalisch-Mathematische Klasse. Sitzungsberichte*, 235–245.

———. (1931). "Zum Kosmologischen Problem der Allgemeinen Relativitätstheorie," *Preussische Akademie der Wissenschaften (Berlin). Physikalisch-Mathematische Klasse. Sitzungsberichte*, 235–237.

———. (1939). "On a Stationary System with Spherical Symmetry Consisting of Many Gravitating Masses," *Annals of Mathematics*, 40, 922–936.

———. (1941). "Demonstration of the Non-Existence of Gravitational Fields with a Non-Vanishing Total Mass Free of Singularities," *Tucumán Universidad Nacional Revisita*, Series A, 2, 11–15.

———. (1949a). "Autobiographical Notes," in P. A. Schilpp (ed), *Albert Einstein: Philosopher–Scientist*. New York: Harper and Row, Volume 1, 1–95.

———. (1949b). "Remarks Concerning the Essays Brought Together in This Co-Operative Volume," in P. A. Schilpp (ed), *Albert Einstein: Philosopher–Scientist*. New York: Harper and Row, Volume 2, 665–688.

———. (1954). *Ideas and Opinions*. New York: Crown.

———. (1955). *The Meaning of Relativity*. 5th ed. Princeton: Princeton University Press.

———. (1961). *Relativity, The Special and General Theory*. New York: Crown Books. First published in 1922 as *Über die Spezielle und die Allgemeine Relativitätstheorie, Gemeinverstandich*. Braunschweig: Vieweg.

———, and Grommer, J. (1927). "Allgemeine Relativitätstheorie und Bewegungsgesetz," *Preussische Akademie der Wissenschaften (Berlin). Physikalisch-Mathematische Klasse. Sitzungsberichte*, 2–13.

———, and Infeld, L. (1949). "On the Motion of Particles in General Relativity Theory," *Canadian Journal of Mathematics*, 1, 209–241.

———, and Pauli, W. (1943). "On the Non-Existence of Regular Stationary Solutions of Relativistic Field Equations," *Annals of Mathematics*, 44, 131–137.

———, and Rosen, N. (1935). "The Particle Problem in the General Theory of Relativity," *Physical Review*, 48, 73–77.

———, and Rosen, N. (1936). "Two-Body Problem in General Relativity Theory," *Physical Review*, 49, 404–405.

———, Infeld, L., and Hoffmann, B. (1938). "The Gravitational Equations and the Problem of Motion," *Annals of Mathematics*, 39, 65–100.

Eisenstaedt, J. (1989). "The Early Interpretation of the Schwarzschild Solution," in D. Howard and J. Stachel (eds), *Einstein and the History of General Relativity, Einstein Studies*, Volume 1. Boston: Birkhäuser, 213–233.

———. (1993). "Lemaître and the Schwarzschild Solution," in J. Earman, M. Janssen, and J. D. Norton (eds), *The Attraction of Gravitation: New Studies in the History of General Relativity, Einstein Studies*, Volume 5. Boston: Birkhäuser, 353–389.

Ellis, G., and Tavakol, R. (1994). "Mixing Properties of Compact $K = -1$ FLRW Models," in D. Hobhill, A. Burd, and A. Coley (eds), *Deterministic Chaos In General Relativity*. New York: Plenum Press, 237–250.

Ellis, G. F. R. (1971). "Topology and Cosmology," *General Relativity and Gravitation*, 2, 7–21.

———. (1988). "Does Inflation Necessarily Imply $\Omega = 1$?" *Classical and Quantum Gravity*, 5, 891–901.

———, and Schmidt, B. G. (1977). "Singular Space-Times," *General Relativity and Gravitation*, 8, 915–953.

———, and Schreiber, G. (1986). "Observational and Dynamical Properties of Small Universes," *Physics Letters A*, 115, 97–107.

———, and Sciama, D. W. (1972). "Global and Non-Global Problems in Cosmology," in L. O'Raifeartaigh (ed), *General Relativity: Papers in Honour of J. L. Synge*. Oxford: Clarendon Press, 35–59.

———, and Stoeger, W. (1988). "Horizons in Inflationary Universes," *Classical and Quantum Gravity*, 5, 207–220.

———, Maartens, R., and Nel, S. D. (1978). "The Expansion of the Universe," *Monthly Notices of the Royal Astronomical Society*, 184, 439–465.

Finkelstein, D. (1958). "Past-Future Asymmetry of the Gravitational Field of a Point Particle," *Physical Review*, 110, 965–967.

Finlay-Freundlich, E. (1951). "Cosmology," in *International Encyclopedia Of Unified Science*, Volume I, No. 8. Chicago: University of Chicago Press.

Fischer, A. E., and Marsden, J. E. (1979). "The Initial Value Problem and the Dynamical Formulation of General Relativity," in S. W. Hawking and W. Israel (eds), *General Relativity: An Einstein Centenary Survey*. Cambridge: Cambridge University Press, 138–211.

Fleming, G. N. (1989). "Lorentz Invariant State Reduction and Localization," in A. Fine and J. Leplin (eds), *PSA 1988*, Vol. 2. East Lansing, MI: Philosophy of Science Association, 112–126.

Fock, V. (1939). "Sur la Mouvement des Masses Finies d'Après la Théorie de Gravitation Einsteinienne," *Academie of Sciences URSS, Journal of Physics*, 1, 81–116.

Forster, M. (1986). "Unification and Scientific Realism Revisited," in A. Fine and P. Machamer (eds), *PSA 1986*, Vol. 1. East Lansing, MI: Philosophy of Science Association, 394–405.

Friedman, J. L., and Morris, M. S. (1991a). "The Cauchy Problem for the Scalar Wave Equation Is Well Defined on a Class of Spacetimes with Closed Timelike Curves," *Physical Review Letters*, 66, 401–404.

———, and Morris, M. S. (1991b). "The Cauchy Problem on Spacetimes with Closed Timelike Curves," *Annals of the New York Academy of Sciences*, 631, 173–181.

———, Morris, M. S., Novikov, I. D., Echeverria, F., Klinkhammer, G., Thorne, K. S., and Yurtsever, U. (1990). "Cauchy Problem in Spacetimes with Closed Timelike Curves," *Physical Review D*, 42, 1915–1930.

———, Papastamatiou, N. J., and Simon, J. Z. (1992). "Failure of Unitarity for Interacting Fields on Spacetimes with Closed Timelike Curves," *Physical Review D*, 46, 4456–4469.

Frolov, V. P. (1991). "Vacuum Polarization in Locally Static Multiply Connected Spacetime and Time-Machine Problem," *Physical Review D*, 43, 3878–3894.
———, and Novikov, I. D. (1990). "Physical Effects in Wormholes and Time Machines," *Physical Review D*, 42, 1057–1065.
Gamov, G. (1970). *My World Line*. New York: Viking Press.
Gardner, M. (1988). "Time Travel," in *Time Travel and Other Mathematical Bewilderments*. New York: W. H. Freeman, 1–14.
Gautreau, R., and Anderson, J. L. (1967). "Directional Singularities in Weyl Gravitational Fields," *Physics Letters A*, 25, 291–292.
Geroch, R. P. (1967). "Topology in General Relativity," *Journal of Mathematical Physics*, 8, 782–786.
———. (1968a). "What is a Singularity in General Relativity?" *Annals of Physics*, 48, 526–540.
———. (1968b). "Local Characterization of Singularities in General Relativity," *Journal of Mathematical Physics*, 9, 450–465.
———. (1969). "Limits of Spacetimes," *Communications in Mathematical Physics*, 13, 180–193.
———. (1970a). "Singularities," in M. Carmeli, S. I. Fickler, and L. Witten (eds), *Relativity*. New York: Plenum Press, 259–291.
———. (1970b). "Domain of Dependence," *Journal of Mathematical Physics*, 11, 437–449.
———. (1977). "Prediction in General Relativity," in J. Earman, C. Glymour, and J. Stachel (eds), *Foundations of Space-Time Theories, Minnesota Studies in the Philosophy of Science*, Volume VIII. Minneapolis: University of Minnesota Press, 81–93.
———, and Horowitz, G. T. (1979). "Global Structure of Spacetimes," in S. W. Hawking and W. Israel (eds), *General Relativity: An Einstein Centenary Survey*. Cambridge: Cambridge University Press, 212–293.
———, and Jang, P. S. (1975). "Motion of a Body in General Relativity," *Journal of Mathematical Physics*, 16, 65–67.
———, Liang, C., and Wald, R. M. (1982). "Singular Boundaries of Space-Times," *Journal of Mathematical Physics*, 23, 432–435.
———, and Traschen, J. (1987). "Strings and Other Distributional Sources in General Relativity," *Physical Review D*, 36, 1017–1031.
Gibbons, G. W., and Hawking, S. W. (1992). "Kinks and Topology Change," *Physical Review Letters*, 69, 1719–1721.
———, Hawking, S. W., and Siklos, S. T. C. (eds) (1983). *The Very Early Universe*. Cambridge: Cambridge University Press.
Glymour, C. (1977). "Indistinguishable Space-Times and the Fundamental Group," in J. Earman, C. Glymour, and J. Stachel (eds), *Foundations of Space-Time Theories, Minnesota Studies in the Philosophy of Science*, Volume VIII. Minneapolis: University of Minnesota Press, 50–60.
Gödel, K. (1949a). "An Example of a New Type of Cosmological Solutions of Einstein's Field Equations of Gravitation," *Reviews of Modern Physics*, 21, 447–450.
———. (1949b). "A Remark about the Relationship between Relativity Theory and Idealistic Philosophy," in P. A. Schilpp (ed), *Albert Einstein: Philosopher–Scientist*. New York: Harper and Row, Vol. 2, 557–562.
———. (1952). "Rotating Universes in General Relativity Theory," *Proceedings of the International Congress of Mathematicians* (1950), 1. Providence: American Mathematical Society, 175–181.

Goldwirth, D. S., Perry, M. J., and Piran, T. (1993). "The Breakdown of Quantum Mechanics in the Presence of Time Machines," *General Relativity and Gravitation*, 25, 7–13.
Gott, J. R. (1991). "Closed Timelike Curves Produced by Pairs of Moving Cosmic Strings: Exact Solutions," *Physical Review Letters*, 66, 1126–1129.
Gowdy, R. H. (1977). "Instantaneous Cauchy Surfaces, Topology Change, and Exploding Black Holes," *Journal of Mathematical Physics*, 18, 1798–1801.
Grant, J. D. E. (1993). "Cosmic Strings and Chronology Protection," *Physical Review D*, 47, 2388–2394.
Grünbaum, A. (1968). *Modern Science and Zeno's Paradoxes*. London: George Allen and Unwin.
———. (1969). "Can an Infinitude of Operations be Performed in a Finite Time?" *British Journal for the Philosophy of Science*, 20, 203–218.
———. (1991). "Creation as a Pseudo-Explanation in Current Physical Cosmology," *Erkenntnis*, 35, 233–254.
Guth, A. H. (1981). "Inflationary Universe: A Possible Solution to the Horizon and Flatness Problems," *Physical Review D*, 23, 347–356.
———, and Steinhardt, P. (1989). "The Inflationary Universe," in P. Davies (ed), *The New Physics*. Cambridge: Cambridge University Press, 34–60.
Hakim, R. (1984). "The Inflationary Universe: A Primer," in Laboratoire "Gravitation et Cosmologie Relativistes," Université Pierre et Marie Curie et C.N.R.S., Institut Henri Poincaré, Paris (ed), *Gravitation, Geometry, and Relativistic Physics, Lecture Notes in Physics*, Vol. 212. Berlin: Springer-Verlag, 302–332.
Harrison, J. (1971). "Dr. Who and the Philosophers: Or Time-Travel for Beginners," *Proceedings of the Aristotelian Society*, Supplementary Volume 45, 1–24.
Hartle, J. B. (1983). "Quantum Cosmology and the Early Universe," in G. W. Gibbons, S. W. Hawking, and S. T. C. Siklos (eds). *The Very Nearly Universe*. Cambridge: Cambridge University Press, 59–89.
Havas, P. (1989). "The Early History of the 'Problem of Motion' in General Relativity," in D. Howard and J. Stachel (eds), *Einstein and the History of General Relativity, Einstein Studies*, Volume 1. Boston: Birkhäuser, 234–276.
———. (1993). "The General Relativistic Two-Body Problem and the Einstein–Silberstein Controversy," in J. Earman, M. Janssen, and J. D. Norton (eds), *The Attraction of Gravitation: New Studies in the History of General Relativity, Einstein Studies*, Volume 5. Boston: Birkhäuser, 88–125.
Hawking, S. W. (1965). "On the Hoyle–Narlikar Theory of Gravitation," *Proceedings of the Royal Society (London) A*, 286, 313–319.
———. (1967). "The Occurrence of Singularities in Cosmology. III. Causality and Singularities," *Proceedings of the Royal Society (London) A*, 300, 187–201.
———. (1979). "Comments on Cosmic Censorship," *General Relativity and Gravitation*, 10, 1047–1049.
———. (1980). "Theoretical Advances in General Relativity," in H. Woolf (ed), *Some Strangeness in the Proportion*. Reading, MA: Addison-Wesley, 145–152.
———. (1988). *A Brief History of Time*. New York: Bantum Books.
———. (1992). "Chronology Protection Conjecture," *Physical Review D*, 46, 603–611.
———, and Ellis, G. F. R. (1973). *The Large Scale Structure of Space-Time*. Cambridge: Cambridge University Press.

———, and Penrose, R. (1970). "The Singularities of Gravitational Collapse and Cosmology," *Proceedings of the Royal Society (London) A*, 314, 529–548.
———, and Stewart, J. M. (1993). "Naked and Thunderbolt Singularities in Black Hole Evaporation," *Nuclear Physics B*, 400, 393–415.
Hilbert, D. (1917). "Die Grundlagen der Physik: Zweiter Mitteilung," *Königliche Gesellschaft der Wissenschaften zu Göttingen. Mathematische-Physikalische Klasse. Nachrichten*, 55–76.
Hochberg, D., and Klephart, T. W. (1994). "Can Semi-Classical Wormholes Solve the Cosmological Horizon Problem?" *General Relativity and Gravitation*, 26, 219–223.
Hogarth, M. L. (1992). "Does General Relativity Allow an Observer to View an Eternity in a Finite Time?" *Foundations of Physics Letters*, 5, 173–181.
———. (1993). "Predicting the Future in Relativistic Spacetimes," *Studies In History and Philosophy of Science*, 24, 721–739.
———. (1994). "Non-Turing Computers and Non-Turing Computability," in D. Hull, M. Forbes, and R. M. Burian (eds), *PSA 1994*, Vol. 1. East Lansing: Philosophy of Science Association, 126–138.
Horowitz, G. T. (1979). "Finding a Statement of Cosmic Censorship," *General Relativity and Gravitation*, 10, 1057–1061.
Horwich, P. (1989). "Time Travel," in *Asymmetries in Time*. Cambridge, MA: MIT Press, 111–128.
Hospers, J. (1967). *An Introduction to Philosophical Analysis*. 2nd ed. Englewood-Cliffs: Prentice-Hall.
Hübner, P., and Ehlers, J. (1991). "Inflation in Curved Model Universes with Non-Critical Density," *Classical and Quantum Gravity*, 8, 333–346.
Israel, W. (1984). "Does a Cosmic Censor Exist?" *Foundations of Physics*, 14, 1049–1059.
———. (1986). "The Formation of Black Holes in Nonspherical Collapse and Cosmic Censorship," *Canadian Journal of Physics*, 64, 120–127.
Janis, A. I., Newman, E. T., and Winicour, J. (1968). "Reality of the Schwarzschild Singularity," *Physical Review Letters*, 20, 878–880.
Jeans, J. (1936). "Man and the Universe," in J. Jeans, W. Bragg, E. V. Appleton, E. Mellanby, J. B. S. Haldane, and J. Huxley (eds), *Scientific Progress*. New York: Macmillan, 13–38.
Johnson, R. A. (1977). "The Bundle Boundary in Some Special Cases," *Journal of Mathematical Physics*, 18, 898–902.
Joshi, P. S. (1985). "Topological Properties of Certain Physically Significant Subsets of Spacetimes," in N. Dadhich, J. Krishna Rao, J. V. Narliker, and C. V. Vishveshwara (eds), *A Random Walk in Relativity and Cosmology*. New York: John Wiley, 128–136.
———. (1993). *Global Aspects in Gravitation and Cosmology*. Oxford: Clarendon Press.
———, and Dwivedi, I. H. (1992). "Naked Singularities in Non-Self-Similar Gravitational Collapse of Radiation Shells," *Physical Review D*, 45, 2147–2150.
———, and Saraykar, R. V. (1987). "Cosmic Censorship and Topology Change in General Relativity," *Physics Letters A*, 120, 111–114.
Kahn, C., and Kahn, F. (1975). "Letters from Einstein to de Sitter on the Nature of the Universe," *Nature*, 257, 451–454.
Kerszberg, P. (1989). *The Invented Universe*. Oxford: Clarendon Press.
Kim, S.-W., and Thorne, K. S. (1991). "Do Vacuum Fluctuations Prevent the Creation of Closed Timelike Curves?" *Physical Review D*, 43, 3929–3947.

King, A. R. (1974). "New Types of Singularity in General Relativity: The General Cylindrically Symmetric Stationary Dust Solution," *Communications in Mathematical Physics*, 38, 157–171.
Klein, F. (1918a). "Bermerkungen über die Beziehungen des De Sitter'schen Koordinatensystem B zu der Allgemeinen Welt Konstanter Positiver Krümmung," *Koninklijke Akademie van Wetenschappen te Amsterdam. Proceedings*, 21 (1918–1919), 614–615.
———. (1918b). "Über die Integralform der Erhaltungssatze und die Theorie der Raumlich-Geschlossen Welt," *Königliche Gesellschaft der Wissenschaften zu Gottingen. Nachrichten*, 394–423.
Klinkhammer, G. (1992). "Vacuum Polarization of Scalar and Spinor Fields near Closed Null Geodesics," *Physical Review D*, 46, 3388–3394.
———, and Thorne, K. S. (1990). "Billiard Balls in Wormhole Spacetimes with Closed Timelike Curves. II. Quantum Theory," preprint.
Kodama, H. (1979). "Inevitability of a Naked Singularity Associated with the Black Hole Evaporation," *Progress in Theoretical Physics*, 62, 1434–1435.
Kostelecký, V. A., and Perry, M. (1994). "No More Spacetime Singularities?" *General Relativity and Gravitation*, 26, 7–12.
Kramer, D., Stephani, H., Herlt, E., and MacCallum, M. (1980). *Exact Solutions of Einstein's Field Equations*. Cambridge: Cambridge University Press.
Kriele, M. (1989). "The Structure of Chronology Violating Sets with Compact Closure," *Classical and Quantum Gravity*, 6, 1607–1611.
———. (1990a). "A Generalization of the Singularity Theorem of Hawking and Penrose to Space-Times with Causality Violations," *Proceedings of the Royal Society (London) A*, 431, 451–464.
———. (1990b). "Causality Violations and Singularities," *General Relativity and Gravitation*, 22, 619–623.
Kristian, J., and Sachs, R. K. (1966). "Observations in Cosmology," *Astrophysical Journal*, 143, 379–399.
Krolak, A. (1986). "Towards a Proof of the Cosmic Censorship Hypothesis," *Classical and Quantum Gravity*, 3, 267–280.
———. (1987a). "Towards a Proof of the Cosmic Censorship Hypothesis in Cosmological Space-Times," *Journal of Mathematical Physics*, 28, 138–141.
———. (1987b). "Strong Cosmic Censorship and the Strong Curvature Singularities," *Journal of Mathematical Physics*, 28, 2685–2687.
Kruskal, M. D. (1960). "Maximal Extension of Schwarzschild Metric," *Physical Review*, 119, 1743–1745.
Kuroda, Y. (1984). "Naked Singularities in the Vaidya Space-Time," *Progress of Theoretical Physics*, 72, 63–72.
Lake, K. (1988). "Comment on 'Naked Singularities in Self-Similar Spherical Gravitational Collapse'," *Physical Review Letters*, 60, 241.
———, and Hellaby, C. (1981). "Collapse of Radiating Fluid Spheres," *Physical Review D*, 24, 3019–3022.
Lanczos, C. (1922a). "Ein Vereinfaschendes Koordinatensystem für die Einsteinschen Gravitationsgleichungen," *Physikalische Zeitschrift*, 23, 537–539.
———. (1922b). "Bemerkung zur de Sitterschen Welt," *Physikalische Zeitschrift*, 23, 539–543.
Landau, L. (1987). "On the Violation of Bell's Inequalities in Quantum Theory," *Physics Letters A*, 120, 54–56.

———, and Lifshitz, E. M. (1962). *The Classical Theory of Fields*. 2nd English ed. Reading, MA: Addision-Wesley.

Lemaître, G. (1932). "L'Univers en Expansion," *Publication du Laboratoire d'Astronomie et de Géodésie de l'Université de Louvain*, 9, 171–205. Also in *Société Scientifique de Bruxelles. Annales A*, 53, 51–85.

Lewis, D. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

———. (1975). "Causation," in E. Sosa (ed), *Causation and Conditionals*. Oxford: Oxford University Press, 180–191.

———. (1986). "The Paradoxes of Time Travel," in *Philosophical Papers*, Volume 2. Oxford: Oxford University Press, 67–80.

Lichnerowicz, A. (1939). *Sur Certains Problèmes Globaux Relatifs au Système des Equations d'Einstein*. Paris: Hermann.

Lifshitz, E. M., and Khalatnikov, I. M. (1963). "Investigations in Relativistic Cosmology," *Advances in Physics*, 12, 185–249.

Lightman, A., and Gingerich, O. (1991). "When Do Anomalies Begin?" *Science*, 255, 690–695.

Linde, A. D. (1980). "Gauge Theories, Time-Dependence of the Gravitational Constant and Antigravity in the Early Universe," *Physics Letters B*, 93, 394–396.

———. (1982). "A New Inflationary Scenario: A Possible Solution of the Horizon, Flatness, Homogeneity, Isotropy, and Primordial Monopole Problems," *Physics Letters B*, 108, 389–393.

———. (1984). "The Inflationary Universe," *Reports on Progress in Physics*, 47, 925–986.

Lossev, A., and Novikov, I. D. (1992). "The Jinn of the Time Machine: Non-Trivial Self-Consistent Solutions," *Classical and Quantum Gravity*, 9, 2309–2321.

Lowe, D. L. (1993). "Semiclassical Approach to Black Hole Evaporation," *Physical Review D*, 47, 2446–2453.

MacBeath, M. (1982). "Who Was Dr. Who's Father?" *Synthese*, 51, 397–430.

MacCallum, M. A. H. (1971). "Mixmaster Universe Problem," *Nature*, 230, 112–113.

———. (1979). "Anisotropic and Inhomogeneous Relativistic Cosmologies," in S. W. Hawking and W. Israel (eds), *General Relativity: An Einstein Centenary Survey*. Cambridge: Cambridge University Press, 533–580.

MacMillan, D. R. (1961). "Cartesian Products on Contractible Open Manifolds," *Bulletin of the American Mathematical Society*, 67, 51–514.

Malament, D. (1977). "Observationally Indistinguishable Space-Times," in J. Earman, C. Glymour, and J. Stachel (eds), *Foundations of Space-Time Theories, Minnesota Studies in the Philosophy of Science*, Volume VIII. Minneapolis: University of Minnesota Press, 61–80.

———. (1985). "Minimal Acccleration Requirements for 'Time Travel' in Gödel Space-Time," *Journal of Mathematical Physics*, 26, 774–777.

———. (1987). "A Note about Closed Timelike Curves in Gödel Space-Time," *Journal of Mathematical Physics*, 28, 2427–2430.

———. (1988). Private communications.

———. (1995). "Introductory Note for *1949b*," in S. Feferman et al. (eds), *Kurt Gödel, Collected Works*, Volume III. New York: Oxford University Press, 261–269.

Maudlin, T. (1990). "Time-Travel and Topology," in A. Fine, M. Forbes, and L. Wessels (eds), *PSA 1990*, Vol. 1. East Lansing, MI: Philosophy of Science Association, 303–315.

Mazenko, G. F., Unruh, W. G., and Wald, R. M. (1985). "Does a Phase Transition in the Early Universe Produce the Conditions Needed for Inflation?" *Physical Review D*, 31, 273–282.

Mellor, D. H. (1981). "Prediction, Time Travel and Backward Causation," in *Real Time*. Cambridge: Cambridge University Press, 160–187.

Menotti, P., and Seminara, D. (1993). "Closed Timelike Curves and the Energy Condition in 2 + 1 Dimensional Gravity," *Physics Letters B*, 301, 25–28.

Michalski, H., and Wainwright, J. (1975). "Killing Vector Fields and the Einstein–Maxwell Field Equations in General Relativity," *General Relativity and Gravitation*, 6, 289–318.

Mikheeva, E. V., and Novikov, I. D. (1992). "Inelastic Billiard Ball in a Spacetime with a Time Machine," preprint.

Mill, J. S. (1843). *A System of Logic, Ratiocinative and Inductive*. London: John W. Parker.

Misner, C. (1963). "The Flatter Regions of Newman, Unti, and Tamburino's Generalized Schwarzschild Space," *Journal of Mathematical Physics*, 4, 924–937.

———. (1967). "Taub–NUT Space as a Counterexample to Almost Anything," in J. Ehlers (ed), *Relativity Theory and Astrophysics: I. Relativity and Cosmology, Lectures in Applied Mathematics*, Volume VIII. Providence: American Mathematical Society, 160–169.

———. (1968). "The Isotropy of the Universe," *Astrophysical Journal*, 151, 431–457.

———. (1969). "Mixmaster Universe," *Physical Review Letters*, 22, 1071–1074.

———, Thorne, K. S., and Wheeler, J. A. (1973). *Gravitation*. San Francisco: W. H. Freeman.

Moncrief, V. (1981). "Infinite-Dimensional Family of Vacuum Cosmological Models with Taub–NUT (Newman–Unti–Tamburino)-type Extensions," *Physical Review D*, 23, 312–315.

———, and Isenberg, J. (1983). "Symmetries of Cosmological Horizons," *Communications in Mathematical Physics*, 89, 387–413.

Morris, M. S., Thorne, K. S., and Yurtsever, U. (1988). "Wormholes, Time Machines, and the Weak Energy Condition," *Physical Review Letters*, 61, 1446–1449.

Newman, R. P. A. C. (1984a). "A Theorem of Cosmic Censorship: A Necessary and Sufficient Condition for Future Asymptotic Predictability," *General Relativity and Gravitation*, 16, 175–192.

———. (1984b). "Cosmic Censorship and Conformal Transformations," *General Relativity and Gravitation*, 16, 943–953.

———. (1986). "Cosmic Censorship and the Strengths of Singularities," in P. G. Bergmann and V. de Sabbata (eds), *Topological Properties and Global Structure of Space-Time*. New York: Plenum Press, 153–168.

———. (1989). "Black Holes Without Singularities," *General Relativity and Gravitation*, 21, 981–995.

———, and Clarke, C. J. S. (1987). "An $\mathbb{R}^4$ Spacetime with a Cauchy Surface Which is Not $\mathbb{R}^3$," *Classical and Quantum Gravity*, 4, 53–60.

Norton, J. D. (1987). "Einstein, the Hole Argument and the Reality of Space," in J. Forge (ed), *Measurement, Realism, and Objectivity*. Dordrecht: D. Reidel, 153–188.

Novikov, I. D. (1989). "An Analysis of the Operation of a Time Machine," *Soviet Journal of Experimental and Theoretical Physics*, 68, 439–443.

———. (1992). "Time Machine and Self-Consistent Evolution in Problems with Self-Interaction," *Physical Review D*, 45, 1989–1994.

Oppenheimer, J. R., and Snyder, H. (1939). "On Continued Gravitational Contraction," *Physical Review*, 56, 455–459.

———, and Volkoff, G. M. (1939). "On Massive Neutron Cores," *Physical Review*, 55, 374–381.
Ori, A. (1991a). "Rapidly Moving Cosmic Strings and Chronology Protection," *Physical Review D*, 44, 2214–2215.
———. (1991b). "Inner Structure of a Charged Black Hole: An Exact Mass-Inflation Solution," *Physical Review Letters*, 67, 789–792.
———. (1993) "Must Time-Machine Construction Violate the Weak Energy Condition?" *Physical Review Letters*, 71, 2517–2520.
———, and Piran, T. (1987). "Naked Singularities in Self-Similar Spherical Gravitational Collapse," *Physical Review Letters*, 59, 2137–2140.
———, and Piran, T. (1988). "Self-Similar Spherical Gravitational Collapse and the Cosmic Censorship Hypothesis," *General Relativity and Gravitation*, 20, 7–13.
———, and Piran, T. (1990). "Naked Singularities and Other Features of Self-Similar General-Relativistic Gravitational Collapse," *Physical Review D*, 42, 1068–1090.
———, and Soen, Y. (1994). "Causality Violation and the Weak Energy Condition," *Physical Review D*, 49, 3990–3997.
Ozsvath, I. (1967). "Homogeneous Lichnerowicz Universes," *Journal of Mathematical Physics*, 8, 326–344.
Padmanabhan, T., and Seshadri, T. R. (1987). "Horizon Problem and Inflation," *Journal of Astrosphysics and Astronomy*, 8, 275–280.
———, and Seshadri, T. R. (1988). "Does Inflation Solve the Horizon Problem?" *Classical and Quantum Gravity*, 5, 221–224.
Pais, A. (1982). *'Subtle is the Lord . . .' The Science and Life of Albert Einstein*. Oxford: Clarendon Press.
Papapetrou, A. (1985). "Formation of a Singularity and Causality," in N. Dadhich, J. Krishna Rao, J. V. Narlikar, and C. V. Vishveshwara (eds), *A Random Walk in Relativity and Cosmology*. New York: John Wiley, 184–191.
———, and Hamoui, A. (1967). "Surfaces Caustiques Degénérées dans la Solution de Tolman. La Singularité Physique en Relativité Générale," *Annales de Institut Henri Poincaré A*, 6, 343–364.
Parker, L., and Fulling, S. A. (1973). "Quantized Matter Fields and the Avoidance of Singularities in General Relativity," *Physical Review D*, 7, 2357–2374.
Patzelt, H. (1990). "On Horizons in Homogeneous Isotropic Universes," *Classical and Quantum Gravity*, 7, 2081–2087.
Penrose, R. (1964). "Conformal Treatment of Infinity," in C. De Witt and B. De Witt (eds), *Relativity, Groups and Topology*. New York: Gordon and Breach, 565–584.
———. (1965). "A Remarkable Property of Plane Waves in General Relativity," *Reviews of Modern Physics*, 37, 215–220.
———. (1969). "Gravitational Collapse: The Role of General Relativity," *Revisita del Nuovo Cimento*, Serie I, 1, Numero Speciale, 252–276.
———. (1973). "Naked Singularities," *Annals of the New York Academy of Sciences*, 224, 125–134.
———. (1974). "Gravitational Collapse," in C. DeWitt (ed), *Gravitational Radiation and Gravitational Collapse*. Dordrecht: D. Reidel, 82–91.
———. (1977). "Space-Time Singularities," in R. Ruffini (ed), *Proceedings of the First Marcel Grossmann Meeting on General Relativity*. Amsterdam: North-Holland, 173–181.

———. (1978). "Singularities of Space-Time," in N. R. Lebovitz, W. H. Reid., and P. O. Vandervoort (eds), *Theoretical Principles in Astrophysics and Relativity*. Chicago: University of Chicago Press, 217–243.
———. (1979). "Singularities and Time-Asymmetry," in S. W. Hawking and W. Israel (eds), *General Relativity: An Einstein Centenary Survey*. Cambridge: Cambridge University Press, 581–638.
———. (1986). Book review of G. W. Gibbons, S. W. Hawking, and S. T. C. Siklos (eds), *The Very Nearly Universe*, in *Observatory*, 106, 20–21.
———. (1988). "Fundamental Asymmetry in Physical Laws," *Proceedings of Symposia in Pure Mathematics*, 48, 317–328.
———. (1989a). *The Emperor's New Mind*. Oxford: Oxford University Press.
———. (1989b). "Difficulties with Inflationary Cosmology," *Annals of the New York Academy of Sciences*, 271, 249–264.
Pfarr, J. (1981). "Time Travel and Gödel's Space," *General Relativity and Gravitation*, 13, 1073–1091.
Pitowsky, I. (1990). "The Physical Church Thesis and Physical Computational Complexity," *Iyyun*, 39, 81–99.
Politzer, H. D. (1992). "Simple Quantum Systems in Spacetimes with Closed Timelike Curves," *Physical Review D*, 46, 4470–4476.
———. (1994). "Path Integrals, Density Matrices, and Information Flow with Closed Timelike Curves," *Physical Review D*, 49, 3981–3989.
Pollock, M. D. (1981). "On the Solution to the Cosmological Horizon Problem Proposed by Zee," *Physical Review D*, 24, 1045–1048.
Quinn, P. (1993). "Creation, Conservation, and the Big Bang," in J. Earman, A. I. Janis, G. Massey, and N. Rescher (eds), *Philosophical Problems of the Internal and External Worlds: Essays Concerning the Philosophy of Adolf Grünbaum*. Pittsburgh: University of Pittsburgh Press, 589–612.
Raine, D. J. (1981). "Mach's Principle and Space-Time Structure," *Reports on Progress in Physics*, 44, 1151–1195.
Ramsey, F. P. (1928–29). "Law and Causality," in D. H. Mellor (ed), *Foundations: Essays in Philosophy, Logic, Mathematics, and Economics* (1978). Atlantic Highlands: Humanities Press, 128–151.
Raychaudhuri, A. K., and Modak, B. (1988). "Cosmological Inflation with Arbitrary Initial Conditions," *Classical and Quantum Gravity*, 5, 225–232.
Rees, M. J. (1972). "Origin of the Cosmic Microwave Background Radiation in a Chaotic Universe," *Physical Review Letters*, 28, 1669–1671.
Reichenbach, H. (1958). *The Philosophy of Space and Time*. New York: Dover.
———. (1971). *The Direction of Time*. Berkeley: University of California Press.
Rendall, A. D. (1992a). "The Initial Value Problem for a Class of General Relativistic Fluid Bodies," *Journal of Mathematical Physics*, 33, 1047–1053.
———. (1992b). "Cosmic Censorship and the Vlasov Equation," *Classical and Quantum Gravity*, 9, L99–L104.
Rindler, W. (1956). "Visual Horizons in World-Models," *Monthly Notices of the Royal Astronomical Society*, 116, 662–677.
Ryan, M. P., and Shepley, L. C. (1975). *Homogeneous Relativistic Cosmologies*. Princeton: Princeton University Press.
Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Sato, H. (1980). "Horizon of the Universe and the Broken-Symmetric Theory of Gravity," *Progress of Theoretical Physics*, 64, 1498–1499.

Savitt, S. (1994). "The Replacement of Time," *Australasian Journal of Philosophy*, 72, 463–474.

Schilpp, P. A. (ed) (1949). *Albert Einstein: Philosopher-Scientist*. Two volumes. New York: Harper and Row.

Schmidt, B. (1971). "A New Definition of Singular Points in General Relativity," *General Relativity and Gravitation*, 1, 269–280.

Schoen, R., and Yau, S.-T. (1983). "The Existence of a Black Hole due to Condensation of Matter," *Communications in Mathematical Physics*, 90, 575–579.

Schwarzschild, K. (1916). "Über das Gravitationfeld eines Massenpunktes nach der Einsteinschen Theorie," *Königlich Preussische Akademie der Wissenschaften (Berlin). Sitzungsberichte*, 189–196.

Sciama. D. W. (1963). "Retarded Potentials and the Expansion of the Universe," *Proceedings of the Royal Society (London), A*, 273, 484–495.

Scott, S. M. (1992). "When Is a Pseudo-Riemannian Manifold Non-Singular?" in Z. Perjés (ed), *Relativity Today*. Commach, NY: Nova Science Publishers, 165–194.

———, and Szekeres, P. (1986a). "The Curzon Singularity. I: Spatial Sections," *General Relativity and Gravitation*, 18, 557–570.

———, and Szekeres, P. (1986b). "The Curzon Singularity. II. Global Picture," *General Relativity and Gravitation*, 18, 571–583.

Seifert, H. J. (1979). "Naked Singularities and Cosmic Censorship: Comment on the Current Situation," *General Relativity and Gravitation*, 10, 1065–1067.

———. (1983). "Black-Holes, Singularities, and Topology," in E. Schmutzer (ed), *Proceedings of the 9th International Conference on General Relativity and Gravitation*. Cambridge: Cambridge University Press, 133–147.

Shapiro, S. L., and Teukolsky, S. A. (1991a). "Formation of Naked Singularities: The Violation of Cosmic Censorship," *Physical Review Letters*, 66, 994–997.

———, and Teukolsky, S. A. (1991b). "Black Holes, Naked Singularities and Cosmic Censorship," *American Scientist*, 79, 330–343.

Siklos, S. T. C. (1979). "Singularities and Invariants," *General Relativity and Gravitation*, 10, 1003–1004.

———. (1981). "Nonscalar Singularities in Spatially Homogeneous Cosmologies," *General Relativity and Gravitation*, 13, 433–441.

Silberstein, L. (1936). "Two-Centers Solution of the Gravitational Equations, and the Need for a Reformed Theory of Matter," *Physical Review*, 49, 268–270.

Smith, J. W. (1986). "Time Travel and Backward Causation," in *Reason, Science and Paradox*. London: Croom Helm, 49–58.

Smoller, J. A. and Wasserman, A. G. (1993). "Existence of Infintely Many Smooth, Static, Global Solutions of the Einstein/Yang–Mills Equations," *Communications in Mathematical Physics*, 151, 303–325.

———, Wasserman, A. G., Yau, S.-T., and McLeod, J.B. (1991). "Smooth Static Solutions of the Einstein/Yang–Mills Equations," *Communications in Mathematical Physics*, 143, 115–147.

Sober, E. (1988). "The Principle of Common Cause," in J. H. Fetzer (ed), *Probability and Causality*. Dordrecht: D. Reidel, 211–228.

———, and Barrett, M. (1992). "Conjunctive Forks and Temporally Asymmetric Inference," *Australasian Journal of Philosophy*, 70, 1–23.

Stachel, J. (1968). "Structure of the Curzon Metric," *Physics Letters A*, 27, 60–61.

———. (1979). "The Genesis of General Relativity," in H. Nelkowski, A. Hermann, H. Posner, R. Schrader, and R. Seiler (eds), *Einstein Symposium Berlin, Lecture Notes in Physics*, Volume 100. Berlin: Springer-Verlag, 428–442.

———. (1986). "Einstein and the Quantum: Fifty Years of Struggle," in R. G. Colodny (ed), *From Quarks to Quasars*. Pittsburgh: University of Pittsburgh Press, 349–385.

———. (1993). "Lanczos' Early Contributions to General Relativity," preprint.

Stein, H. (1970). "On the Paradoxical Time-Structures of Gödel," *Philosophy of Science*, 37, 589–601.

———. (1990). "Introductory Note to *1949a*," in S. Feferman, J. W. Dawson, Jr., S. C. Kleene, G. H. Moore, R. M. Solovay, and J. van Heijenoort (eds), *Kurt Gödel, Collected Works*, Volume II. New York: Oxford University Press, 199–201.

———. (1995). "Introductory Note to *1946/9*," in S. Feferman et al. (eds), *Kurt Gödel, Collected Works*, Volume III. New York: Oxford University Press, 202–229.

Steinmuller, B., King, A. R., and Lasota, J. P. (1975). "Radiating Bodies and Naked Singularities," *Physics Letters A*, 51, 191–192.

Stewart, J. M. (1968). "Neutrino Viscosity in Cosmological Models," *Astrophysical Letters*, 2, 133–135.

Summers, S. J., and Werner, R. (1987a). "Bell's Inequalities and Quantum Field Theory. I. General Setting," *Journal of Mathematical Physics*, 28, 2440–2447.

———, and Werner, R. (1987b). "Bell's Inequalities and Quantum Field Theory. II. Bell's Inequalities are Maximally Violated in the Vacuum," *Journal of Mathematical Physics*, 28, 2448–2456.

Susmann, R. A. (1988). "On Spherically Symmetric Shear-Free Perfect Fluid Configurations (Neutral and Charged). II. Equation of State and Singularities," *Journal of Mathematical Physics*, 29, 945–970.

Swinburne, R. (1968). *Space and Time*. London: Macmillan.

Synge, J. L. (1950). "The Gravitational Field of a Particle," *Proceedings of the Royal Irish Academy A*, 53, 83–114.

———. (1960). *Relativity: The General Theory*. Amsterdam: North-Holland.

Szekeres, G. (1960). "On the Singularities of a Riemannian Manifold," *Publicationes Mathematicae, Debrecen*, 7, 285–301.

Thom, P. (1975). "Time-Travel and Non-Fatal Suicide," *Philosophical Studies*, 27, 211–216.

Thomson, J. F. (1954–55). "Tasks and Super-Tasks," *Analysis*, XV, 1–13.

't Hooft, G. (1992). "Causality in (2 + 1)-dimensional Gravity," *Classical and Quantum Gravity*, 9, 1335–1348.

Thorne, K. S. (1991). "Do The Laws of Physics Permit Closed Timelike Curves?" *Annals of the New York Academy of Sciences*, 631, 182–193.

———. (1994). *Black Holes and Time Warps: Einstein's Outrageous Legacy*. New York: W. W. Norton.

Tipler, F. J. (1974). "Rotating Cylinders and the Possibility of Global Causality Violation," *Physical Review D*, 9, 2203–2206.

———. (1976). "Causality Violation in Asymptotically Flat Space-Times," *Physical Review Letters*, 37, 879–882.

———. (1977a). "Singularities and Causality Violation," *Annals of Physics*, 108, 1–36.
———. (1977b). "Singularities in Conformally Flat Space-Times," *Physics Letters A*, 64, 8–10.
———. (1978). "Energy Conditions and Spacetime Singularities," *Physical Review D*, 17, 2521–2528.
———. (1979). "What Is a Black Hole?" *General Relativity and Gravitation*, 10, 1063–1067.
———. (1980). "General Relativity and the Eternal Return," in F. J. Tipler (ed), *Essays in General Relativity*. New York: Academic Press, 21–37.
———. (1985). "Note on Cosmic Censorship," *General Relativity and Gravitation*, 17, 499–507.
———, Clarke, C. J. S., and Ellis, G. F. R. (1980). "Singularities and Horizons—A Review Article," in A. Held (ed), *General Relativity and Gravitation*, Volume 2. New York: Plenum Press, 97–206.
Tolman, R. C. (1931a). " On the Problem of the Entropy of the Universe as a Whole," *Physical Review*, 37, 1639–1660.
———. (1931b). "On the Theoretical Requirements for a Periodic Behaviour of the Universe," *Physical Review*, 38, 1758–1771.
———, and Ward, M. (1932). "On the Behavior of Non-Static Models of the Universe when the Cosmological Term Is Omitted," *Physical Review*, 39, 835–843.
Torretti, R. (1979). "Mathematical Theories and Philosophical Insights in Cosmology," in H. Nelkowski, A. Hermann, H. Posner, R. Schrader, and R. Seiler (eds), *Einstein Symposium Berlin, Lecture Notes in Physics*, Volume 100. Berlin: Springer-Verlag, 320–335.
———. (1983). *Relativity and Geometry*. New York: Pergamon.
Turner, M. S. (1987). "Inflation in the Universe, Circa 1986," in M. A. H. MacCallum (ed), *General Relativity and Gravitation: Proceedings of the 11th International Conference on General Relativity and Gravitation*. Cambridge: Cambridge University Press, 223–246.
van Bendegam, J. P. (1994). "Ross's Paradox is an Impossible Super-Task," *British Journal for the Philosophy of Science*. Forthcoming.
van Fraassen, B. C. (1980). *The Scientific Image*. Oxford: Clarendon Press.
———. (1989). *Laws and Symmetry*. Oxford: Clarendon Press.
van Stockum, W. J. (1937). "The Gravitational Field of a Distribution of Particles Rotating about an Axis of Symmetry," *Proceedings of the Royal Society of Edinburgh*, 57, 135–154.
Vickers, J. A. G. (1987). "Generalized Cosmic Strings," *Classical and Quantum Gravity*, 4, 1–9.
Visser, M. (1993). "From Wormhole to Time Machine: Remarks on Hawking's Chronology Protection Conjecture," *Physical Review D*, 47, 554–565.
———. (1994). "Van Vleck Determinants: Traversable Wormhole Spacetimes," *Physical Review D*, 49, 3963–3980.
Wald, R. M. (1973). "On Perturbations of a Kerr Black Hole," *Journal of Mathematical Physics*, 14, 1453–1461.
———. (1980). "Dynamics in Nonglobally Hyperbolic, Static Space-Times," *Physical Review*, 21, 2802–2805.
———. (1984a). *General Relativity*. Chicago: University of Chicago Press.

———. (1984b). "Black Holes, Singularities and Predictability," in S. M. Christenson (ed), *Quantum Theory of Gravity*. Bristol: Adam Hilger, 160–168.
———. (1986). "Inflation and Phase Transitions," in E. W. Kolb, M. S. Turner, D. Lindley, K. Olive, and D. Seckel (eds), *Inner Space/Outer Space*. Chicago: University of Chicago Press, 341–344.
———. (1992). "Correlations Beyond the Cosmological Horizon," to appear in the proceedings of the workshop "Origin of Structure in the Universe," Point d'Oyre, Belgium, 1992.
———, and Iyer, V. (1991). "Trapped Surfaces in the Schwarzschild Geometry and Cosmic Censorship," *Physical Review D*, 44, 3719–3722.
———, and Yurtsever, U. (1991). "General Proof of the Averaged Null Energy Condition for a Massless Scalar Field in Two-Dimensional Curved Spacetime," *Physical Review D*, 44, 403–416.
Weir, S. (1988). "Closed Time and Causal Loops: A Defence Against Mellor," *Analysis*, 48, 203–209.
Wells, H. G. (1968). *Time Machine*. New York: Bantam Books.
Weyl, H. (1917). "Zur Gravitationstheorie," *Annalen der Physik*, 54, 117–145.
———. (1919). *Raum–Zeit–Materie*. 3rd ed. Berlin: Julius Springer.
Wheeler, J. A., and Feynman, R. P. (1949). "Classical Electrodynamics in Terms of Direct Interparticle Action," *Reviews of Modern Physics*, 21, 425–433.
Will, C. M. (1993). *Theory and Experiment in Gravitational Physics*. Revised edition. Cambridge: Cambridge University Press.
Yodzis, P. Seifert, H.-J., and Müller zum Hagen, H. (1973). "On the Occurrence of Naked Singularities in General Relativity," *Communications in Mathematical Physics*, 34, 135–148.
Yourgrau, P. (1991). *The Disappearance of Time*. Cambridge: Cambridge University Press.
Yurtsever, U. (1990). "Test Fields on Compact Space-Times," *Journal of Mathematical Physics*, 31, 3064–3078.
———. (1991). "Classical and Quantum Instability of Compact Cauchy Horizons in Two Dimensions," *Classical and Quantum Gravity*, 8, 1127–1139.
Zee, A. (1980). "Horizon Problem and the Broken-Symmetric Theory of Gravity," *Physical Review Letters*, 44, 703–706.
Zemach, E. M. (1968). "Many Times," *Analysis*, 28, 145–151.

Index

(Italicized page numbers denote definitions)