The Generation of Knowledge through Experimentation in Fundamental Physics: The Case of Gravity Probe B

Christopher EVANS

Universitat Autònoma de Barcelona

TESI DOCTORAL

The Generation of Knowledge through Experimentation in Fundamental Physics: The Case of Gravity Probe B

Autor:
Christopher EVANS

Signat: _____

Director de Tesi:
Dr. H Carl HOEFER

Signat: _____

Tutor:
Dr. Thomas STURM

Aquesta tesi es ofereix d'acord amb els requeriments del

Doctorat en Filosofia

del *Departament de Filosofia* de la *Facultat de Filosofia i Lletres*

Setembre de 2016

Certificado de Dirección

Título de la tesis doctoral:

The Generation of Knowledge through Experimentation in Fundamental Physics: The Case of Gravity Probe B

Director de la tesis:

Dr. H Carl Hoefer

Doctorando:

Christopher Evans

Programa de doctorado: Doctorado en Filosofía
del Departamento de Filosofía
de la Facultad de Filosofía y Letras
Universitat Autònoma de Barcelona

Visto bueno del director de la tesis.

Firmado en Barcelona, el 26 de septiembre de 2016:

_____

H Carl Hoefer

UNIVERSITAT AUTÒNOMA DE BARCELONA

Resumen

Facultat de Filosofia i Lletres

Departament de Filosofia

Doctorat en Filosofia

La Generación de Conocimiento a través de la Experimentación en la Física Fundamental: El Caso de "Gravity Probe B"

escrito por Christopher EVANS

En el presente trabajo analizo desde un punto de vista crítico el episodio que representa “Gravity Probe B” (GP-B) (La Sonda de la Gravedad B) en la historia de la experimentación en la física fundamental. Anunciado como una “Prueba del Universo de Einstein”, GP-B fue un experimento para examinar predicciones de la teoría general de la relatividad (TGR) que duró 50 años. GP-B nació en Stanford University, en el momento que el principio de la tecnología de satélites lo convirtió en una posibilidad y se convirtió en el proyecto más longevo de la NASA: los resultados finales se publicaron en el año 2011. Siguiendo el diseño original de 1960, GP-B pretendió medir el arrastramiento del marco espaciotemporal y el efecto geodésico sobre un giroscopio en órbita alrededor de la tierra en un satélite funcionando en modo “drag-free” (sin resistencia). Siguiendo una órbita puramente gravitacional, junto con el giroscopio y los magnetómetros formados por dispositivos superconductores de interferencia cuántica, la nave espacial contenía un telescopio que rastreaba una estrella guía como punto de referencia. La misión espacial empezó en el 2004 y concluyó en el 2005 con el objetivo de medir el cambio en la orientación del eje de giro de los giroscopios, relativo al inmóvil espacio inercial, con una precisión de 0.5 milésimas de un segundo de arco ($\sim 10^{-7}$

grados) a lo largo de un año. Para realizar el experimento fue necesario desarrollar varias tecnologías completamente novedosas, y los sistemas de abordo diseñados establecieron varios récords por ser los sistemas más cerca de la perfección diseñados jamás. (GP-B) representa una oportunidad única para analizar cómo funcionan los experimentos científicos extremos y una gran oportunidad de estudiar los esfuerzos para generar conocimiento acerca de la TRG basado en la experimentación. GP-B se encontró con serias dificultades durante la ejecución, con importantes anomalías y un ruido excesivo en los datos. El equipo se vio obligado a desarrollar controvertidos métodos nuevos para analizar los datos que se obtuvieron. Inicialmente presento tanto la física relevante a la aproximación específica a la TRG apropiada para analizar la gravitación en el sistema solar (el marco posnewtoniano de parametrización) como la historia de la confirmación de TRG. Después de presentar GP-B y sus objetivos, introduzco el marco analítico que adopto para examinar los resultados y conclusiones que el equipo logró. Utilizo el trabajo de James Woodward y Deborah Mayo, combinándolo en una perspectiva basada en tres puntos: los datos observados pueden ser evidencia para fenómenos teóricos subyacentes; la experimentación hace lo posible para rastrear la veracidad de las hipótesis a través de la sensibilidad contrafáctica de los datos a las afirmaciones teóricas; y para que los datos valgan como evidencia a favor de un fenómeno, la prueba que represente el encaje de estos con las predicciones de las hipótesis examinadas debe ser severo, aunque no necesariamente represente un uso novedoso de los datos. Destaco muchas preocupaciones con el análisis de los datos producidos por GP-B, pero a través de mi análisis basada en este marco, mi conclusión es que las afirmaciones del equipo de GP-B son perfectamente válidas. También indico que este episodio demuestra que puede ser importante que los científicos adopten las perspectivas más sofisticadas propuestas por filósofos en lugar de contar con los más comunes acercamientos epistemológicos. Finalmente, indico que a pesar de la posibilidad de que el conocimiento generado no sea del todo sólido e inmóvil, y que algún día pueda revisarse, cumple con los requisitos más estrictos que la sociedad normalmente pide de las conclusiones de la investigación.

UNIVERSITAT AUTÒNOMA DE BARCELONA

Abstract

Facultat de Filosofia i Lletres

Departament de Filosofia

Doctorat en Filosofia

The Generation of Knowledge through Experimentation in Fundamental Physics: The Case of Gravity Probe B

by Christopher EVANS

In this thesis, I critically analyse Gravity Probe B (GP-B) as an extraordinary episode in the history of experimentation in fundamental physics. Billed as "Testing Einstein's Universe," GP-B was a 50-year-long experiment to test crucial predictions of the General Theory of Relativity (GTR). GP-B started life at Stanford University when satellite technology first made the "Relativity Gyroscope Experiment" feasible and it went on to become the longest running mission in NASA's history; final results were published in 2011. Following the original design published in 1960, GP-B set out to measure frame dragging (also known as the Lense-Thirring effect) and the geodetic (or de Sitter) effect on a superconducting gyroscope orbiting the Earth in a "drag-free" satellite. Essentially executing a purely gravitational orbit, together with the science instrument assembly containing the (multiple) gyroscope(s) and superconducting quantum interference devices used as magnetometers, the spacecraft housed a telescope trained on a reference "guide star". The mission flew from 2004 to 2005 and aimed to measure the change in the orientation of the spin axis of the gyroscopes, relative to "fixed" inertial space identified using the guide star, to within 0.5 milliarcseconds ($\sim 10^{-7}$ degrees) over the year-long experiment. The experiment required the development of several

completely new technologies before it could be performed and the on-board systems broke numerous records as the most nearly perfect and most sensitive systems created. It represents a unique opportunity to analyse the workings of scientific experimentation taken to the extreme and a rare chance to examine efforts to generate knowledge based on experimental GTR: one of our two current fundamental physics theories. GP-B encountered serious problems during execution of the space mission with major anomalies and excessive noise in the data collected. The team was forced to develop controversial new data analysis methods to attempt to salvage meaningful results from the unexpected and unrepeatable dataset they retrieved. I initially present both the physics of GTR in the specific weak gravity approximation appropriate for analysing gravitational effects within the Solar System (the parametrised post-Newtonian framework) and the prior history of confirmation of GTR. After presenting GP-B and its aims, I then introduce the analytical framework that I adopt to examine the claims made by the team regarding their data analysis and eventual findings. I draw heavily on work by James Woodward and Deborah Mayo, among others, and combine this into a 3-point approach: observed data can act as evidence for underlying theoretical phenomena; experimentation contrives to track the truth of hypotheses via the counterfactual sensitivity of the data produced by the specific experimental set-up to those theoretical claims; and for data to count as evidence in favour of a phenomenon, the test that the match between them and the predictions of the hypothesis being examined represents must be severe, although not necessarily entail novel use of the data. I highlight many worries with the GP-B data analysis, but through analysing it within this framework, I conclude that the claims of the GP-B team are valid. I also indicate that the episode shows how it can be important for working scientists to adopt the more sophisticated approaches advocated by some philosophers rather than relying on more typical epistemological attitudes found in 20th century textbooks. I close by noting that although the knowledge gained may not be unshakeably solid and is open to future revision, it fulfils the strictest demands normally placed by society on the conclusions of investigation.

Acknowledgements

First and foremost I would like to thank my supervisor, Carl Hoefer, for his limitless help and encouragement, as well as his patience; I don't feel I need to say that without him this project would never have got off the ground and much less have been brought to a satisfying close. Thanks also to Thomas Sturm for stepping in and agreeing to be my tutor over the last year of my work to produce this thesis; but maybe much more importantly, for encouraging me to continue with my research when I had almost abandoned it and always expressing his faith in me. My thanks also go to Adán Sus who taught me an awful lot about general relativity as well as the philosophy of physics in general, during the early stages of my research.

I first presented the research that formed the basis for this thesis at the SEFA 2010 conference in Tenerife; I am very grateful to my audience there for their useful comments and discussion. I would also particularly like to thank those who organised and attended the PSX2 workshop in Konstanz, where I presented a preliminary version of this work; especially Allan Franklin and Deborah Mayo for their helpful comments and discussion as well as encouragement to continue my research.

I would like to thank Clifford Will for permission to reproduce the three diagrams of his that appear here in Chapter 2. I would also like to thank Francis Everitt for sending me and allowing me to reproduce several of the figures, as well as for the interest he expressed in my work. Also my thanks to Jennifer Spencer who was the GP-B Web Site Curator for her help and advice; and to Liz Goheen of the Perimeter Institute for Theoretical Physics, for sending me the egg that appears here as Figure 5.4. I would also like to thank both Dick McCray and Nick Cornish for agreeing to share their thoughts with me regarding the NASA decision to withdraw funding for GP-B, years after they were instrumental in that decision.

I would like to say thank you to my sister Katy for proofreading this for me, making helpful comments regarding the style, and always encouraging me to keep working; all the remaining mistakes are quite definitely my own!

Finally, I have left the most import to last: my everlasting thanks to Jeimy. I am indebted to her for her patience, understanding, encouragement and, over the final few weeks, putting all her architectural projects aside and acting as my research assistant. This work has benefited greatly from her contributions to it, especially her expert assistance with the figures as well as her work on the typesetting and bibliography.

The research for this work was partially funded by the Spanish public through grants awarded by both the Spanish national government, via different ministries over the years: grants FFI2008-06418-C03-03 (2008-2011) and FFI2011-29834-C03-03 (2012-2014); and the Catalan regional authorities, via their financial support of the GRECC Research Group (*Grup de Recerca en Epistemologia i Ciències Cognitives*): award 2009SGR1528 (2009-2013).

Contents

List of Figures

List of Tables

List of Abbreviations

ANOVA	analysis of variance
C-rules	correspondence rules
EEP	Einstein Equivalence Principle
EFE	the Einstein field equations
GP-B	Gravity Probe B
GSV	guide star valid
GTR	General Theory of Relativity
HD	hypothetico-deductive (method or model of scientific practice)
HF	high frequency
KACST	King Abdulaziz City for Science and Technology (Saudi Arabia)
LF	low frequency
LIGO	Laser Interferometer Gravitational Observatory
LLI	local Lorentz invariance
LM	London moment
LPI	local position invariance
mas	milliarcseconds (thousandths of one second of arc: 1 mas $\approx 3 \times 10^{-7}$degrees $\approx 5 \times 10^{-9}$ rad)
NASA	National Aeronautics and Space Administration
NSRC	NASA (Science Mission Directorate, Astrophysics Division) Senior Review Committee
PCA	principal component analysis
PI	principal investigator
PPN	parametrised post-Newtonian
RV	Received View of science
SAC	(GP-B) Science Advisory Committee
SC	(Mayo's) severity criterion

SIA	scientific instrument assembly
SQUID	superconducting quantum interference devices
STR	special theory of relativity
TFM	trapped flux mapping
UGR	universality of the gravitational redshift
UN	use novelty
VLBI	very-long baseline interferometry
WWII	the Second World War (1939-1945)

Chapter 1

Introduction

1.1 Inspiration, Motivation and Progress

In this initial, introductory chapter of my doctoral thesis, I aim to place the work that I detail in the remaining chapters in context. It is important for the introduction to a thesis to set out clearly and unambiguously the research questions that are addressed in the thesis. I will of course do just that in this chapter; but I also aim to do rather more. By way of placing my research within its context, I will explain my reasons for embarking down the specific road or paths that I have followed, and recount some of the twists and turns that have led to or resulted from the decisions that I have made, together with the evolution of both my research and the subject matter that is at the heart of it. While there will be little detailed analysis in this introduction, I trust that in providing a general context for the science that I consider, the specific philosophical questions I aim to answer and my personal motivation, it will thereby naturally lead into the chapters that follow, provide the reader with orientation, create expectations, and also stimulate the reader's interest.

It is the nature of research that it grows and changes as we progress towards our initial objectives and shift both our expectations and targets as we advance. It is often the case that the issues that the actual scientific results obtained or the progression of a research project shed most light on are subtly different from the questions that the researchers originally aimed to address. In some cases, as advances are made and we learn more, research unfolds and develops in wholly unexpected ways. This does not at all mean that as researchers we

are straying from the overall goal of applying our skills and abilities in such a way as to make a meaningful contribution to the growth and improvement of the knowledge base that is commonly held by humans and which we can hope will one day be freely available and accessible to all.[1] It can indeed be argued that the most significant steps forward in the advancement of human knowledge are made not through the practice of what, as an extension of Thomas Kuhn's famous—or maybe these days infamous, given the criticism it has received—division of science back in 1962, I will term problem-solving research (Kuhn, 1996[1962]); but by the extraordinary and unexpected. Be that as it may, any and all research may vary and change as it progresses and initial problem-solving research can lead to unforeseen breakthroughs.

To illustrate the way in which research objectives are refined and altered as our investigation progresses, we need only consider the classical metaphor of knowledge being a clearing in the woods; with ourselves, the knowing subjects, situated in the clearing, and each tree left standing and visible to us around the clearing being an unanswered question. As we increase our knowledge, so we enlarge the clearing; thus, with the increasing perimeter of the clearing, more and more trees—unanswered questions—become visible to us. Although it is quite normal to be able to see parts of some trees behind the front row which stands before us in full sight, and maybe to catch the occasional glimpse of trees further off in the forest, we cannot get a clear view of those trees until we have decided on a direction and felled a path towards them. Many of the trees that we thus discover as we expand our clearing were previously entirely hidden from us. Our interest may be drawn to certain

[1] I would just like to note here in passing that one of the specific aims of the European Union's Horizon 2020 initiative ("not as an end in itself but as a tool to facilitate and improve the circulation of information" the EU Fact Sheet tells us) is to promote open access which it defines as "the practice of providing on-line access to scientific information that is free of charge to the end-user" as a way of accelerating research, avoiding duplication, enhancing interdisciplinarity and other collaborative efforts, and improving returns on investment in research together with increasing accountability which should "ultimately benefit society and citizens". It is in keeping with these aims and objectives that I hope that one day access will be available and free to all. (Fact sheet: Open Access in Horizon 2020 https://ec.europa.eu/programmes/horizon2020/sites/horizon2020/files/FactSheet_Open_Access.pdf accessed on 09/08/2016)

questions or issues (trees) that become visible to us as we continue to expand our knowledge (the clearing) and so we may decide to head towards those next. Alternatively, progress in the direction that we had originally chosen may be halted by insurmountable questions—seemingly unfellable trees—or ones we do not wish to touch once they are in full view or even while we can only partially discern them further off and envisage that they will halt our progress if we continue in the same direction. So it is with research: we may decide which direction to take, which question to address initially and begin to plot a path; but what we will find, and the issues and options that open up to us as we advance were previously hidden from us. The novelty of the perspective we gain may well cause us to alter our direction radically or to pick out a complex path that leads us away from our original course.

If such possibilities are part of the nature of research, then it is almost inevitable that the research that leads to the production of a doctoral thesis will change as the novice researcher explores their own interests, hones their research skills and discovers new and unexpected angles and perspectives; and possibly whole new questions to focus on. The area of research that has led to the case study at the heart of the thesis that you are now reading was first suggested to me by Carl Hoefer back in 2004. At that time I was studying the graduate course "An Introduction to the Philosophy of Physics" given by him for the first time that year at the *Universitat Autònoma de Barcelona*. The GP-B space mission, run jointly by NASA and a team at Stanford University, was readying for launch and it was generating considerable interest within specialist circles as the project team led by C.W.F., "Francis", Everitt (and scrutinised at regular intervals by a Science Advisory Committee (SAC)[2] chaired by the undoubted expert on the subject Clifford Will) prepared to place the GP-B satellite in orbit around the Earth and start the experiment. The space mission consisted of delivering the satellite carrying the experimental equipment into a near-Earth orbit, and then monitoring it and collecting data over a total period of some 18 months. As I read more on GP-B, I came to see it as the momentous project I believe it to have been and I glimpsed the potential it offered not just for the advancement of science, but as an exceptional example of

[2] https://einstein.stanford.edu/MISSION/mission2.html#sac accessed on 09/08/2016

contemporary scientific experimentation within my budding field of expertise. The entire GP-B project was a colossal attempt to bridge the gap between, on the one hand, technological experimental design and the detection of effects so subtle that we were only just beginning to be able to measure them; and, on the other hand, some of our most basic and fundamental theoretical physics. It seemed clear to me that such a case study would offer an unparalleled opportunity to scrutinise the generation of knowledge through the practices of contemporary experimentation in fundamental physics.

Hence, one of my first research objectives was to examine the details of the theoretical framework within which GP-B was to provide us with evidence in favour of—or against—Einstein's General Theory of Relativity (GTR). To do this, I needed to return to my academic origins in physics, and learn more of GTR. Here there were two directions to follow simultaneously: firstly, Einstein's initial theoretical achievements and how they have been added to and adapted since he first announced his theory in November 1915; and secondly, how observation and experimentation had developed to provide evidence in favour of the theory or posed new challenges for it. The philosophy of physics is renowned for being of limited access to academics from other branches of philosophy; it is often claimed that in-depth knowledge of physics is necessary in order to address the philosophical issues at the heart of physics meaningfully. Whether that is true or not—and I have seen many insightful and productive contributions made to discussions on issues in the philosophy of physics from philosophers who are certainly not experts in the field—a solid base is no doubt needed to appreciate many of the subtleties that arise. This is where Chapter 2 and Chapter 3 of this thesis stem from. In the former (*Solar System Tests of Einstein's General Theory of Relativity*; see Section 1.2 below), I hope to provide the non-physicist reader with just such a base, so that they will have a context within which to place GP-B and its goals. In the latter (*Gravity Probe B: An Experiment in General Relativity*; see Section 1.3), I examine the working of GP-B in more detail. I report the difficulties initially facing the project; I analyse just how the project team designed the experiment to overcome those difficulties; and I consider both the technical requirements they recognised as necessary for the success of the experiment and the possible results they expected from the execution of the experiment

via the specific apparatus and set-up they had designed and produced.

In tandem with my research into GP-B, and more strictly within the realm of the philosophy of science, I examined and considered the different approaches that have been adopted in the analysis of scientific theories and practice during the lifetime of GTR; and even how they may have been influenced by it. It is often said that one of the functions or roles of philosophy (and particularly of philosophy of science) is to provide checks and balances on other activities: to bring to light and scrutinise the assumptions that may be hidden and what is implicit in different claims; and then to assess their validity and effects. Moreover, an additional goal may be to identify practices within the processes that we believe lead to the production of knowledge that we can expect to produce repeatable results and knowledge that can usefully and predictably be applied in situations other than those in which it was discovered. We could certainly claim that this is one of the longest running and most fundamental issues in the whole of the area of modern philosophy of science. In more general terms, as Gary Cutting tells us when discussing the three-pronged approach to revealing the nature of scientific methodology—strictly philosophical, often via metaphysics and epistemology; internal debate among scientists, particularly in times of major scientific upheaval; and more recently, that of historians who study the practice of science:

> After the triumph of Newtonian science, philosophical reflection on methodology (from Locke on) has been in the very different position of starting from the unquestionable success of a scientific paradigm. The question is no longer how to build an engine of scientific progress, but how to understand and justify the one we have.
>
> (Cutting, 2001, p. 426)

This was the road I set out along with my sights fixed on analysing the whole GP-B project, which as I say, I saw—and continue to see—as not just potentially a remarkable advance in science, but more importantly for my purposes, as a unique opportunity to study scientific experimentation, specifically within the area of fundamental physics. So, in Chapter 4 (*Counterfactual Difference Makers and Severe Tests*; see Section 1.4 below), I consider briefly the changing climate or fashions over the 20th century within the philosophy of science. I

go on to examine the developments that I believe have led us to our current position from which we examine scientific experimentation as the particular activity it actually is and analyse it both accurately and usefully in the light of what has gone before, but from perspectives that were not previously the norm.

My main concern and the focus of my analysis is a contemporary approach to the analysis of scientific practice and knowledge generation through experimentation akin to what has come to be known as New Experimentalism.[3] Via such an approach, I consider how, together with what we have learned from all the dominant trends within the philosophy of science throughout the 20th century, we can best analyse scientific data as evidence for underlying phenomena. I introduce counterfactual reasoning as a more appropriate way (than those normally adopted in purely syntactic or semantic approaches) to analyse and assess the adequacy of scientific practices from what I consider to be a more naturalised perspective. I also focus on the importance of statistics for the modern experimental scientist; particularly what has been called standard error statistics (Mayo, 1996).

With my knowledge of GTR, the background to the GP-B project and the different approaches to the philosophy of science up to date, I was ready to tackle an analysis of the GP-B results and claims. However, as so often happens in cutting-edge scientific experimentation and research, the results of GP-B were not as expected. It seems to have been clear right from the start of the space mission that the experimental apparatus or set-up was not performing as expected; but for reasons that were far from clear. The space mission had an almost precisely set maximum time to run, and after analysing the initial phase, decisions were made and the appropriate changes implemented to allow the collection of what was hoped would still be meaningful data to continue. The overall GP-B space mission then concluded on 29th September 2005, with the total depletion of the on-board supply of liquid helium and

[3] As I explain in a footnote in Chapter 4 (footnote 7) this term was first adopted, it appears, in Ackermann, 1989, and it has come to be used to refer to the contemporary approach to the philosophy of science that stresses the importance of experimentation and actual laboratory practices instead of placing excessive weight on theories, theorising and what has been called standard analytic epistemology.

the GP-B satellite entering silent hibernation; and the task of data analysis commenced in earnest. The team here on Earth was equipped with perfect replicas of much of the apparatus and this allowed the members of the team to continue experimenting, during and after the mission, to try to understand the anomalies in the data and what had caused them.

In this way, GP-B entered a new and unexpected phase; and in tandem, so did my research. Through years of analysis of the data, the team adapted existing theory and developed novel techniques and models to explain the results, refine them and thereby rescue the project and produce what they claimed were meaningful, accurate and precise results that pushed the observational limits on GTR to new extremes. Not everyone was convinced by their methods, however, and after over-running by several years, NASA dropped out of the project and left it for the Stanford team to continue with a new partner: the King Abdulaziz City of Science and Technology in Saudi Arabia. After publishing their "Post Flight Analysis: Final Report" in March 2007 and "Science Results—NASA Final Report" in December 2008, just after the decision by NASA not to fund the continuing data analysis efforts any further, the team eventually announced their final results in May 2011. In Chapter 5 (*Gravity Probe B Science Results*; see Section 1.5) I analyse the most important results of GP-B and explain how they varied from expectations, as well as the different ways in which the team worked around the limitations imposed by the noisy nature of their primary read-out. Just as the GP-B team had switched the emphasis of their efforts, so my analysis shifted to what had become the central question of the validity of their workaround and the refined, reformulated new results. I moved into new areas, following the trajectory of the GP-B team. I had already become familiar with the work of the philosopher and statistician Deborah Mayo, and here I drew on her expertise in data analysis and the formation of genuine knowledge from it, as a resource with which to consider and assess the methodology of the GP-B team.

So, in Chapter 6, (*Lessons to be Learned*; see Section 1.6) I offer my analysis of the work of the GP-B team, from the perspective of New Experimentalism (laid out in some detail in previous chapters). Thus, Chapter 5 and Chapter 6

contain the most important novel analytical work of mine in this thesis, drawing on all the previous chapters. It is followed by the very brief Chapter 7 where I make some closing remarks and consider future research. In that final brief chapter, I also consider directions that future research may take, leading on from the work presented here.

I can now state the overall research question that I address through the whole of the work that you have before you, as follows:

> Can we consider the GP-B project to be an example of the generation of genuine scientific knowledge through experimentation in terms of the warrant that we can assign to the team's findings?

I divide this global question into three closely related issues.

- Can we expect the knowledge that the GP-B team claim to have arrived at to be solid, incontrovertible and permanent; or should we see it as provisional and revisable?
- Should we see the GP-B team as having subjected their claims to sufficiently stringent tests and can we therefore see the data that they collected and analysed over so many years as useful evidence in favour of the underlying phenomena of gravitation that GTR predicts?
- Should our broader non-expert society accept the findings of the GP-B team as rigorous and a novel contribution to human knowledge?

Of course, to arrive at answers to these, I will have to answer a whole host of preliminary questions concerning GTR and its confirmation or otherwise, the GP-B set-up and mission, how we assess and judge knowledge claims in experimental science, the status of knowledge itself, and the actual process of data analysis that the GP-B team undertook to draw their conclusions. That work, which is precisely what the whole of this doctoral thesis contains, is what I will now briefly introduce in the following sections of this chapter.

1.2 The Physics: GTR in the Solar System

In Chapter 2, by means of an introduction for the reader who is not familiar with the subject, I first track Einstein's development of GTR from the particular case of inertial motion that he had developed in his prior Special Theory of Relativity (STR), via the universality of free fall or the Principle of Equivalence, as it is also known. I explain how Einstein extended this equivalence from the strictly mechanical case that Galileo had expressed, to a more general requirement and thus arrived at the characteristic of general covariance for GTR that he considered to be central to the theory. I mention how the extension that Einstein introduced to the Principle of Equivalence was the same as the combination of local Lorentz invariance (LLI) and local position invariance (LPI), which are properties introduced into all metric theories of spacetime, and also how evidence for STR amounts to evidence for the adequateness of the metric formulation of spacetime.

To introduce how GTR can be compared and contrasted with other metric theories of spacetime within a low-speed, weak-field scenario, such as is afforded by conditions within the Solar System, I then move on to explain the basis of post-Newtonian parametrisation. I explain how the parameters introduced within this formalism are used to compare different metric theories and where GTR sits in the spectrum of possible theories that can be distinguished within the framework. This allows me to present the three classic tests of GTR put forward by Einstein himself in 1916: the observed anomalous motion of mercury, the deflection of light by a massive body and the gravitational redshift. The first of these three tests has been objected to by detractors who point to the possible importance of the non-novel nature of the test, as the anomaly was very well known to Einstein before he formulated GTR, and its persistence must be seen as motivation to reconsider gravitation, even if it was never one of his specific aims to account for it. (In fact, given the confidence that Einstein expressed in his finished theory, it seems quite clear that he would have gone ahead and announced its completion and have been convinced that in GTR he had developed the lacking theory of gravitation whether it had accounted for the Mercurian anomaly or not!) Nevertheless, within the theoretical framework adopted here (expressed mathematically as

a combination of both the relevant post-Newtonian parameters) the observed motion of Mercury sets limits on the overall deviation from the results yielded by GTR of approximately 1 part in 10^3. The second test sets a limit on one of the 2 post-Newtonian parameters to within 1 part in 10^4; though again, not without controversy. This has since been refined to 1 part in 10^5 through the adoption of modern time-delay techniques that were unknown to Einstein. Finally, the third of Einstein's tests must be seen as a test of all metric theories and not specifically of GTR. Since all the serious contenders as rival theories of gravitation are also metric theories which must by definition obey the same equivalence as that which Einstein insists on (and indeed as he did as early as 1907) and this is what this effect tests, the gravitational redshift cannot be used to distinguish between GTR and other metric theories of gravitation, but must be seen as supporting the metric nature of the theory as the most appropriate way to represent gravitation.

Having thus established the state of play, in terms of the evidence for GTR prior to GP-B, I then consider the two effects the project was conceived, designed, built and executed to measure; and thereby extend our knowledge of the match between the predictions of GTR and observation. These were frame dragging, the detailed calculation of which as it can be derived from the equations of GTR was published by 1918 by Lense and Thirring; and the de Sitter (or geodetic) effect, which had already been calculated and published two years earlier in 1916 (the same year as Einstein published his initial summary of GTR[4]). Both of these effects result from the distortion of the local spacetime around a massive spinning body and can be detected through the interaction between such a body and a second spinning test body orbiting the first. As I go on to emphasise, frame dragging (sometimes referred to as a gravitomagnetic effect, due to parallels that can be drawn with the electromagnetic case) is a tiny effect that results specifically from the spin components of the motion of the two bodies and is in no way even contemplated within the framework of Newtonian mechanics. It is therefore sometimes erroneously seen as a more qualitative test of the departure from

[4] As I explain in a footnote in Chapter 2 (footnote 2) although Einstein announced his GTR in 1915 in a brief communication, it was not until 1916 that he published a detailed summary of it.

Newtonian mechanics, whereas in fact neither effect can be accounted for in any way within that framework. Frame dragging is, however, much the smaller of the two effects for a spinning body such as the Earth and therefore requires a much more sensitive experiment to observe it. For the specific set-up and characteristics of GP-B (my case study), the deviation in the orientation of a spinning test body orbiting the Earth due to this effect is calculated to be: 0.039 arcseconds per year (the duration of the data acquisition phase of GP-B was set to be of the order of one year, so this was an indication of the precision required). The geodetic effect predicted by GTR is two orders of magnitude larger: a total geodetic effect of 6.6 arcseconds per year. As I then go on to explain in the following chapter (Chapter 3), the experimental design was to allow the signals from these two effects to be separated out, as they would be perpendicular to each other.

Having thus established the theoretical background, I next move on to consider the particular case of GP-B that I followed as it unfolded and now analyse here in this dissertation.

1.3 The Experiment: GP-B

Following on from the considerations outlined above, I next provide—in Chapter 3—the details necessary for the uninitiated reader to be able to appreciate the experimental design of GP-B and the execution of the space mission. As I indicate from the start of this work, the mission did not go as planned. The initial results that it provided made a major change of emphasis necessary together with the adoption of novel data analysis techniques. Moreover, many additional years of work were required (and they seem to have been strongly criticised and questioned) to eventually produce final results. However, before tackling those issues, it is important to understand the experiment as it was planned, designed and executed; the results that were expected from it; and the entire process of experimentally testing GTR that it embodied.

I fill in some historical details of the inception of the project at the end of the 1950s and suggest that the decision to go ahead with such an enormously

complicated and costly project may well have been more politically motivated than scientifically. The project emerged under funding from the U.S. Air Force and Department of Defense, prior to NASA joining the incipient project in 1964: the period of the Cold War when the space race was more hotly contended than ever. However, through its long history of over 50 years, from the initial idea being published in 1960 to the announcement of the final results in 2011, and considering the obstacles it has overcome, there can be no doubt that above all GP-B was, in many different aspects, an enormously important scientific research project and physics experiment.

In describing the experiment, it is important to have its goals well established: the measuring of the geodetic effect, published by de Sitter in 1916, and the frame-dragging effect, the calculation of which was published by Lense and Thirring in 1918. As the latter of these had never been measured before the year of the launch of GP-B (in contrast to the former), it was often considered to be the principal goal of the experiment. To explain the geodetic effect, I describe in some detail the idea of parallel transport: how curved space can result in angular shifts in the orientation of vectors that are parallel transported around a closed circuit in that curved space. This is in essence the effect of transporting a free-falling gyroscope around a geodesic of warped spacetime. That is exactly what GP-B was designed to do; together with measuring the cumulative effect of that shift in orientation, over thousands of such circuits, during the (roughly year-long) experiment. As I explain, the GP-B spacecraft containing a gyroscope (or in fact multiple gyroscopes) was placed in a near-Earth orbit. It was then left in free fall with the gyro spinning inside, also in free fall, and then the orientation of the gyro spin axis was measured to establish how it changed over the course of the experiment. I briefly describe the experimental arrangement and technologies that were employed to achieve this prior to the successful launch of the space mission on 20th April 2004.

I then go on to discuss the ways in which the GP-B team, over the course of the entire project, time and again managed to adapt to the circumstances they were confronted by, or manoeuvre things to their advantage in one way or another. I mention here the worries that inevitably arise from accumulating ever

more and more complex levels of modelling when it comes to assessing any scientific experiment. As a long-running cutting-edge and extremely complex experiment, GP-B was forced to take on board many assumptions and adopt many models devised in different (however closely related) fields and areas. In such circumstances, the dangers of misinterpretation and misrepresentation grow to an unquantifiable extent. In this chapter I detail the requirements for the success of the experiment, one of which was the accurate alignment of the on-board telescope with distant fixed stars representing inertial space. As I point out, this relies on Earth-based very-long baseline interferometry (VLBI) techniques; technology that adds an entirely new dimension to our interpretation of the final GP-B results. Furthermore, it may raise the spectre of circularity in our use of GTR to establish the baseline against which GP-B took measurements, when GP-B itself was designed to test GTR; something I return to in Chapter 5.

Another of the most important aspects of the entire GP-B space mission was the development and then the correct functioning of drag-free near-Earth orbit flight. In such an orbit there are always effects on any spacecraft which tend to produce perturbations (albeit tiny ones) in the purely gravitational free-fall orbit. Whereas for many purposes this is not a significant or relevant effect, for GP-B it was absolutely crucial, since to test GTR effectively the gyroscope at the heart of the experiment had to be in a purely gravitational orbit: travelling along a path that can be seen as tracing out a geodesic of the warped spacetime around the Earth. This important advance in satellite technology was partially developed specifically for the GP-B mission; always with much interest from other potential beneficiaries, including the military, which, as I have said, was one of the most important sources of initial interest in the project. Another of the technologies developed during the lifetime of GP-B and essential for ensuring the required precision in the determination of both the gyroscope spin axis and the spacecraft orientations, was that employed in superconducting quantum interference devices (SQUIDs) together with pioneering work on the London moment (LM).

Finally there was the data processing: both that which it was initially planned to use to decipher the GTR traces within the output dataset and the techniques

that the team developed after the space mission in the face of what initially seemed like insurmountable problems. Those problems involved, though were not limited to, the degree of noise in the data and its origin. The team identified the source as trapped magnetic flux within the superconducting coating of the gyroscopes, and set about developing methods to recover the GTR signal from among the noise. I consider this in detail later on, in Chapter 5.

1.4 The Philosophy of Science: Knowledge through Experimentation

In Chapter 4, I consider the mainstream development of the philosophy of science since Einstein first published his GTR a century ago. His theory has been seen as marking a crucial change to our entire world view. Beforehand, we lived in a common-sense, flat Euclidean universe which was readily accessible to us. Einstein showed us that such an appearance is merely an embodiment of human parochialism or short-sightedness compared to when we consider things on astronomical scales and examine the universe outside our Solar System. He opened our eyes to the counterintuitive, irregularly curved space-time that we actually seem to inhabit. GTR was also a triumph of a genuinely physical interpretation of aspects of the pure mathematics developed in the nineteenth century through which he explained an observed anomaly that had previously defied satisfactory explanation: the precession of Mercury's perihelion. This led to the optimistic opinion that through learning how to interpret other aspects of mathematics in similar ways we could eventually learn all the secrets of the universe around us. This was in stark contrast to the view that major scientific questions were totally insoluble, as propounded famously by Emil du Bois-Reymond towards the end of the nineteenth century[5].

[5]For example, Susan Haack in her 2003 book (Haack, 2003, p. 330) cites the two "celebrated" lectures published by du Bois-Reymond in 1872 and 1880 in which he expressed and reiterated such a view.

So it is that GTR can be seen as instrumental in the new attitudes towards and understanding of science offered by logical positivism early in the twentieth century. I briefly review some of the key aspects of this way of interpreting science and scientific theories before considering its shortcomings. These include it being a reconstructionist approach which can only possibly offer an idealised account of what science is and how it works, as opposed to offering a more naturalistic account of the practice of science as we can observe it happening in the laboratory. The untenable strict division it calls for between theoretical and observational vocabularies, and just what is meant by the partial interpretations that it claims to offer also contributed to alternative approaches being sought. I go on to examine the semantic view that came to challenge and even replace the law-sentential view favoured by logical empiricism as the majority view of the best way to describe and explain the practices of science and the theories it produces. As a direct contrast to the syntactic view, as its name indicates, the semantic approach aimed to capture all the meaning embodied in a theory, rather than express it in a unique way. The key to achieving this is the use of models and seeing a theory as some form of a collection of all the models that represent its theoretical content.

I note that GTR can be seen as conforming to aspects of both of these approaches to the analysis of scientific theories very well. However, I consider that they are not the best means by which to analyse the generation of knowledge through scientific experimentation. So I then move to what I consider to be the most promising way to analyse just what it is that scientific observation and experimentation actually does through recording (unpredictable and idiosyncratic) data that can be seen as evidence for underlying (unobservable) phenomenon which are the subject of theoretical interest. As I explain, important aspects of this approach were propounded by Bogen and Woodward (Bogen and Woodward, 1988) as an alternative to the previous efforts to reconstruct scientific practice almost exclusively in terms of logical relationships and models; and instead to consider how scientists actually set about making discoveries, building evidence for the phenomena that theory predicts and generating knowledge.

As opposed to always seeing the reliability of science as stemming from the

relationship of warrant between some supposed, idealised data and theory, Woodward in particular advocates considering the ways in which scientists manipulate apparatus to gain the counterfactual sensitivity necessary to identify genuine difference makers. In an attempt to broaden the options available to philosophers when analysing scientific practice, he breaks the two-step process (one step between theory and prediction, and the other between prediction and—potential—observational confirmation) into three steps: one between background theory and the prediction of phenomena; an intermediate step of experimental design and execution to establish how phenomena can—potentially—be detected through the trace they leave in data; and the matching of actual data collected to prediction. In this way it is claimed that the reliability of science can be located, not just in the logical relationships that hold between background theory and prediction but also in the practical work of teasing out the necessary counterfactual sensitivity in specific experimental set-ups. This seems to reflect accurately the way much experimentation in mature science actually progresses towards knowledge production, but it is lacking one vital ingredient: statistics. This is supplied by the work pioneered by Mayo, which I mention above (pages 6 and 7).

Mayo goes into great detail examining, analysing and explaining just how it is that scientists arrive at stringent tests for their hypotheses. Through setting out what she terms a severity criterion (SC), she also shows that standard epistemological approaches, while successfully capturing certain aspects of the severity of the tests scientific hypotheses are subjected to, fails to analyse this relationship fully. In particular, Mayo shows that novelty (of the use of data), which is often regarded as essential for a test of a hypothesis to be considered as stringent or severe, is just one aspect of this desired property and not a necessary condition for severity at all. Mayo champions the idea that double counting, as such re-use of data has been called, does not necessarily invalidate the claims of evidential support that can be derived from the match between data and prediction. I take Mayo's work and combine it with that of Bogen and Woodward to arrive at what I consider to be the most appropriate approach to the analysis of scientific experimentation for my analysis in the following chapters (Chapter 5 and Chapter 6) of the GP-B results and claims. The approach that I thus build consists of three vital ingredients: recognition of

the separation between theoretical phenomena and data that may contain the vital telltale trace of those phenomena; the extra component, beyond logical prediction, that counterfactual sensitivity in the laboratory requires if genuine difference makers are to be identified through experimental intervention; and the necessary conception of severity if we are to be able to identify reliable knowledge generation.

1.5 GP-B Results: Severely Testing

I cannot hope here to reproduce much of the nearly six years of data analysis that the GP-B team undertook. In Chapter 5, rather I try to bring to the fore the essence of what they did and how they produced their rabbit out of the big black hat that GP-B appeared to have sunk into. Before they could start to look for solutions the team faced at least two vital, make-or-break questions. Could a source—or sources—of the excessive noise that the final dataset contained be established? Although it is often the case that effects that are not fully understood can be accounted for through design and manipulation, in this case, with a unique dataset already generated, analysis and treatment of the noisy data seemed totally impossible without having some hold on the processes at work in the production of that noise. The second question was whether it was possible to identify patterns within the perturbations evident in the data that were sufficiently different from both the motion of the spacecraft and the target relativity trace signal for the interference to be modelled and methods devised to calculate it and subtract it out.

With such important problems looming, it is impossible not to understand NASA's decision to devote its resources to more promising endeavours. But apparently all was not lost. When spending of the order of 1,000 million US dollars on an experiment, it is important to cover as many eventualities as possible. The team had some leeway. The final, post Science Phase of the mission allowed for much "fiddling" with the apparatus. The aim of this had always been to provide calibration data by putting the satellite through exaggerated moves that it had not undergone during the delicate Science Phase and to see how the systems performed compared to its previous

handling during the on-orbit initialisation. As it happened, this provided vital information regarding the nature of the perturbations and allowed the team to confirm that they were electromagnetic in origin.

The project may have been saved thanks to the redundancy built into the systems and specifically the engineering data that had been collected sporadically throughout the entire mission. In order to register how the on-board systems were working in minute detail, this information was much more fine grained than the science data, set to detect the tiniest of shifts over a year. But with a (partial) means of access to this information and an idea of what had been causing the problems was it possible to model the new effects accurately enough to remove them from the actual science data readout and leave a "clean" signal?

Here I apply the considerations of the severity of the tests to which we subject our hypotheses that I present in the previous chapter. By examining the counterfactual relationships that the actual science data stand in to the phenomena hypothesised to be behind them our aim is always to identify the actual difference makers. I consider whether the team was justified in "cleaning up" data to the extent that was required. Alternatively, I assess whether the mere fact that the dataset was compromised meant that the team would never be able to show to a convincing degree that they did not just continue to work on the data until they arrived at the answer they wanted. Certainly that was the worry of many at NASA. But there are times when we should accept that perseverance pays off and the results of a Herculean team effort can exceed all expectations. Certainly, that is what the team claimed and through my analysis of applying the approach I outline in the previous chapter I find no fault in their reasoning or claims.

But scepticism may remain. I conclude the chapter by imagining how the case may be argued one way or another. This leaves me to go on in the next chapter to consider what the status of their claims of novel knowledge really is.

1.6 Aspects of GP-B Knowledge Claims

My analysis of the actions of the GP-B team encourage me to accept their claims; but I just want to consider exactly what that means. Having opted for a naturalistic vision of the justification scientists aim for of the reliability of the knowledge they produce, I now ask what status our knowledge claims can have if we take a path of rejecting logical entailment all the way from theory to data. So I consider the status of the GP-B knowledge claims from a Quinean-like perspective. Quine famously cast reductionism as one of the dogmas that we need to avoid if we are to advance in our understanding of how it is that we can build up genuine knowledge that goes far beyond our immediate experience. The Quinean-like web of knowledge that I introduce here helps to show just how tentative belief that relies on so much secondary and novel knowledge must be. Like all our beliefs, of course this new knowledge is potentially revisable; but in this case, considering the multiple layers of modelling behind the reasoning and the novelty of many of the techniques adopted, even if the reasoning seems sound, we must be tentative in our acceptance.

I then return to Mayo and the analytical scheme I have developed through this work. I spell out the argumentation that I see the GP-B team as having implicitly followed in reaching their claims: I make explicit some of the reasoning behind their claims. I probe the sensitivity that is necessary according to the scheme I have developed for the match between data and prediction to be considered a severe test and therefore to offer evidential support. There are at least two levels on which the GP-B results could be seen to fail the requirement for novelty of use, if we were to consider that to be a necessary condition for the match between data and prediction to constitute a severe test (precisely what Mayo—and I with her—reject). Firstly, on the global level, if we consider the GP-B data to form a unique dataset, and it was certainly all gathered from the SQUID output signals during the experiment, then it seems impossible to deny that the data itself was used to devise the model which it is then claimed it matches, thereby offering evidence in favour of GTR. This is clearly a case of double counting. Secondly, if we look at the specific details of the technique developed to analyse the data from each of

the gyros, the so-called trapped flux mapping, then it is also clearly a case of double counting. But having rejected the use-novelty requirement, it remains to be seen whether despite this non-novel characteristic, the match of data to prediction can still be considered a severe test of the theoretical predictions.

Once again, the weight of evidence seems to be on the side of the GP-B team and their claims to have tested Einstein's universe; but is that enough? That is the question I address in the final section of Chapter 6 when I adopt what could be seen as a more social point of view and consider what broader society may make of the work that I report here. What standard does society accept as proof? Can we draw parallels between this case and things that are closer to the decisions that lay members of society may ordinarily be called on to make?

I certainly think we can; and I see this bridging an important gap between epistemological and social questions. Thus, here I reconstruct a partial trial, with the public as jurors. If society applies its strictest criteria of evidence to the GP-B findings, can the jury find the supposed causes responsible for the observed perturbations and thereby lend credence to the claims of the GP-B team? Or alternatively, should the public throw the case out on the grounds that there is no evidence in favour of the far-fetched hypotheses put forward by the team other than the data they already had; and it would be impossible to convict with no independent corroboration of the initial evidence? Through this unorthodox approach I hope to go some way to demonstrating ways in which the analysis of even the most complex scientific procedures can be brought out of its ivory tower and a little closer to the general public (or at least we can try to find a window in that tower to provide a glimpse of it).

I bring my analysis to a close, with some brief conclusions (Chapter 7). My aim is to have considered different perspectives from which we may consider claims considering the generation of knowledge through experimentation, particularly in the case of general relativity. I believe the case study I present here is illuminating in this respect. I must admit that I was initially surprised by the conclusion I reach that supports the claims of the GP-B team; but of course, that is part of the beauty of research: we never know where it is going to lead us.

Chapter 2

Solar System Tests of Einstein's General Theory of Relativity

"Still, even if you aren't [right], you have the right idea; don't be afraid of looking at the equations just because you can't follow the derivations."

Everitt quoting Bill Fairbank's reply to him when he made an early suggestion concerning the gyroscope experiment (Fairbank et al., 1988, p. 49)

2.1 Introduction: The Lie of the Land

In this chapter, I review the predictions and the confirmation, prior to the launch of the Gravity Probe B (GP-B) satellite in 2004, of Einstein's General Theory of Relativity (GTR) within the solar system. (I deal almost exclusively with Einstein's final version of the theory, as its historical development is not my concern here. I only briefly mention the papers from 1905 to 1915 in which he developed his ideas and continued to alter them before arriving at the finished[1] theory in November 1915.[2]) This serves the double objective of reviewing the state of our knowledge at that time, 2004, and providing a base

[1] Except for the later addition and subsequent removal of the cosmological term.

[2] Einstein's final version of GTR is contained in a series of papers which he sent to the Prussian Academy in November 1915. He then published a review of the theory in 1916 (Einstein, 1952[1916]). The literature contains many references to the theory as being from both 1915 and 1916; but I will follow what seems to be the more common choice, which emphasises the earlier date as the date of conception, and generally refer to it as his 1915 theory. That said, I will often mention his 1916 review of GTR.

from which to examine the aims and expectations of the GP-B experiment; which I consider in Chapter 3.

All Solar System experiments and observations fit a weak-field (and for massive bodies, low-speed) approximation very well. Einstein initially used a method of post-Newtonian approximation to solve the problem of the anomalies observed in the orbit of Mercury.[3] This type of approach has since been extensively developed and is commonly used to analyse Solar System experiments. Within such a framework, "correcting" terms are added to the calculations and predictions of Newtonian physics in order to take account of the effects predicted by GTR (and other competing metric theories of gravity). This guarantees Newtonian gravity as a first approximation and allows us to use the small deviations from it within the Solar System to compare our observations with the predictions of GTR and its rival theories.

I begin the chapter by considering different ideas concerning the principle of equivalence, which is a vital link between classical mechanics and Einstein's relativity and thereby forms a starting point for GTR. Then in Section 2.3, I sketch the parametrised post-Newtonian (PPN) framework which is the specific model that has been developed to compare and contrast weak-field experimental results and observations with theoretical predictions. In Section 2.4, I consider the 3 "classic" tests of GTR proposed by Einstein, which are fundamental to an understanding of the extent to which his theory has been confirmed. Since I go on to look at the GP-B experiment, the next section is dedicated to the de Sitter (or geodetic) effect and the Lense-Thirring effect (also known as frame dragging or the gravitomagnetic effect). These are the two effects predicted by GTR which it was hoped GP-B would measure, thereby refining the limits on acceptable models of gravity and consequently restricting possible rival theories. Finally, in Section 2.6, I look at two more weak-field, low-speed tests of GTR: the Shapiro time delay and the Nordtvedt effect. I end the section by mentioning the effects of strong gravity that are accessible to us.

[3] After explaining how he ensured that his equations gave Newtonian gravity as a first approximation, Einstein tells us of his initial calculations concerning Mercury, prior to the development of the Schwarzschild solution: "To solve this problem I made use of the method of successive approximation." (Einstein, 2005[1922], p. 94).

I consider it to be beyond the scope of this work to go into details concerning the mathematical derivation and analysis of the formalism used here. The work is well established and can be found in many standard texts, particularly Will, 1993a; and although I draw on it heavily, I offer nothing new to the mathematics. Furthermore, I have limited this work to considerations of the weak-field approximation of Solar System experiments and observations. Therefore, I do not consider gravitational waves, as reported for the first time in February 2016 by the LIGO Scientific Collaboration and the Libra Collaboration (Abbott et al., 2016),[4] which were predicted by Einstein in his 1915 GTR. However, it should be noted that just like frame dragging and the geodetic effect, gravitational waves are a phenomenon that was totally novel to GTR and therefore could have no counterpart in Newtonian theory to which a post-Newtonian "perturbation" could be applied. An important difference between the phenomena of both the geodetic effect and frame dragging on the one hand, and gravitational waves on the other is that while the former are predicted by GTR to be produced at detectable levels within the Solar System (by a massive spinning body such as the Earth, for example), the latter is only predicted to occur at detectable levels strictly within the realm of strong gravity and therefore requires colossal, extra-Solar System events. The expanding and fascinating fields of extra-Solar System observational gravity and its highly theoretical counterpart, seen as so vital to the resolution of the apparent contradictions between our two current best fundamental theories of physics, quantum gravity, are areas into which I hope to extend my work in the future.

[4]The report was published on 11th February, but the reported observation was made on 14th September 2015.

2.2 The Principle of Equivalence

2.2.1 The Weak Equivalence Principle

As is well known,[5] after publishing his special theory of relativity (STR) in 1905, Einstein wished to extend the principle of the relativity of motion to include reference frames in arbitrary motion; not just the special case of inertial (non-accelerated) motion. He was particularly concerned with the way in which absolute space was considered to affect ponderable matter by defining inertial frames of reference, but absolute space was not in turn affected by that matter. Through considering Mach's principle, which introduces the idea that the effects of inertia are not relative to an absolute space but rather to the relative position of all gravitating matter, he realised that the key to extending his theory may lie in the existing Galilean, or Newtonian, principle of equivalence. This states that all test bodies[6] fall at the same rate (ignoring fluid resistance) in the same gravitational field, irrespective of their mass or internal structure. In Newton's theory this was just an empirical result, and indeed appeared to be rather a coincidence; we should surely expect that the mass or internal structure of an object will affect how it reacts to an external force that pulls it down, as it were.

The principle is often known as the universality of free fall (UFF).[7] It is a way of saying that a body's passive gravitational mass (a measure of the degree to which the body is affected by gravity) must be directly proportional to its inertial mass (which determines its resistance to being accelerated). If

[5] See, for example, Norton, 2005, p. 5, where he tells us that, "Einstein speculates immediately on the possibility of extending the principle of relativity to accelerated motion. He suggests the relevance of gravitation to this possibility and posits what is later called the principle of equivalence as the first step towards the complete extension of the principle of relativity."

[6] "Test" bodies here are theoretical models that do not experience tidal forces due to their spatial extension within the gravitational field, nor do they cause perturbations in the local gravitational field. They can be considered as idealisations of well-defined pieces of matter in the limit as their mass tends to zero.

[7] In the work that is often referred to as *The Bible* when it comes to gravitation, Misner, Thorne, and Wheeler, 1973, p. 348, the authors refer to this as the "uniqueness of free fall".

this equivalence did not hold true, bodies having different ratios of passive gravitational mass to inertial mass would behave differently in the same gravitational field, and this has never been observed to be the case. Newtonian mechanics offers no explanation for this equivalence and simply accepts it as an empirically observed coincidence. In typical fashion, bringing his philosophical understanding and insight to bear on the physics, this appeared unsatisfactory to Einstein:

> ... classical mechanics offers no explanation for this equality. It is however, clear that science is fully justified is assigning such a numerical equality only after this numerical equality is reduced to an equality of the real nature of the two concepts.
>
> (Einstein, 2005[1922], pp. 56-7)

He therefore set about searching for a new model of the interaction between matter and spacetime which included the equivalence of passive and gravitational mass. This was eventually to lead to the revolution of seeing gravitation not as some kind of force that acts differently on each body according its properties; but rather as a type of topography or geometry that obliges all bodies, irrespective of their structure, to behave in a similar way. For this reason, the principle of equivalence is of primary importance to the whole of the later development of gravitation theory and GTR. Indeed it has become so inseparable that some tests that were previously thought to be tests of GTR have since been shown to test only the more fundamental principle of equivalence, which other metric theories of gravity must also obey. So, following Einstein's footprints historically, this is where we should start to consider just what relativity really is and what it means. We can thus see the equivalence principle as both motivating Einstein's search for a theory of gravity and as being the link that binds his revolutionary work to what went before; it enables us to see his revolution as the evolution of what went before.

To quantify the limits of the equivalence between inertial mass and passive gravitational mass we can assume that the former (m_I) differs from the latter (m_P) by gravity interacting not only with the inertial mass, but also with some additional internal energy (E^A) resulting from an interaction "A", then:

$$m_P = m_I + \sum_A \frac{\eta^A E^A}{c^2} \tag{2.1}$$

where η^A is a dimensionless constant which depends on the contribution the internal energy of interaction "A" makes to the passive gravitational mass. The interaction of this passive mass with local gravity produces an acceleration (a) on the body. By comparing the accelerations of two different bodies (for our purposes, with different compositions so that the amount of additional internal energy is different) resulting from the same local gravity, a ratio of accelerations can be obtained which is independent of gravity. Experiments designed to measure these accelerations then lead to the so-called Eötvös ratio, given by:

$$\eta \equiv \frac{2\,|a_1 - a_2|}{|a_1 + a_2|} = \sum_A \eta^A \left(\frac{E_1^A}{m_1 c^2} - \frac{E_2^A}{m_2 c^2} \right) \tag{2.2}$$

where m_1 and m_2 are the inertial masses of the two bodies.

Experiments have placed limits on deviations from strict equivalence of passive gravitational mass and inertial mass ever since Galileo's original work and his (thought) experiment involving the tower at Pisa. Like the ratio given in 2.2, such experiments are often known as Eötvös(-type) experiments after Baron von Eötvös, who obtained exceptionally accurate experimental results (about 1 part in 10^8) using a torsion balance, in a series of experiments he performed between 1888 and 1908. It seems that Einstein became aware of the work of Eötvös in 1907, and he explicitly cites it in a footnote in his 1916 review of GTR. However, if this is the case, Eötvös's work certainly could not have motivated Einstein's desire to extend STR in this way before 1907. Figure 2.1 shows the evolution of the limit on this parameter resulting from different experiments prior to the launch of GP-B in 2004;[8] experimental evidence limits

[8] In general the best values from experimental results I use in this chapter are from Will, 2006 (unless I explicitly state otherwise), from around the time of the GP-B experiment which lasted from 2004 to 2005. However, Will has since updated that publication to include later experiments, but none has substantially improved the best limit, which he cites as having decreased from $\eta < 3 \times 10^{-13}$ in 2006 to $\eta < 2 \times 10^{-13}$ in 2014, both from the "Eöt-Wash" experiments performed at the

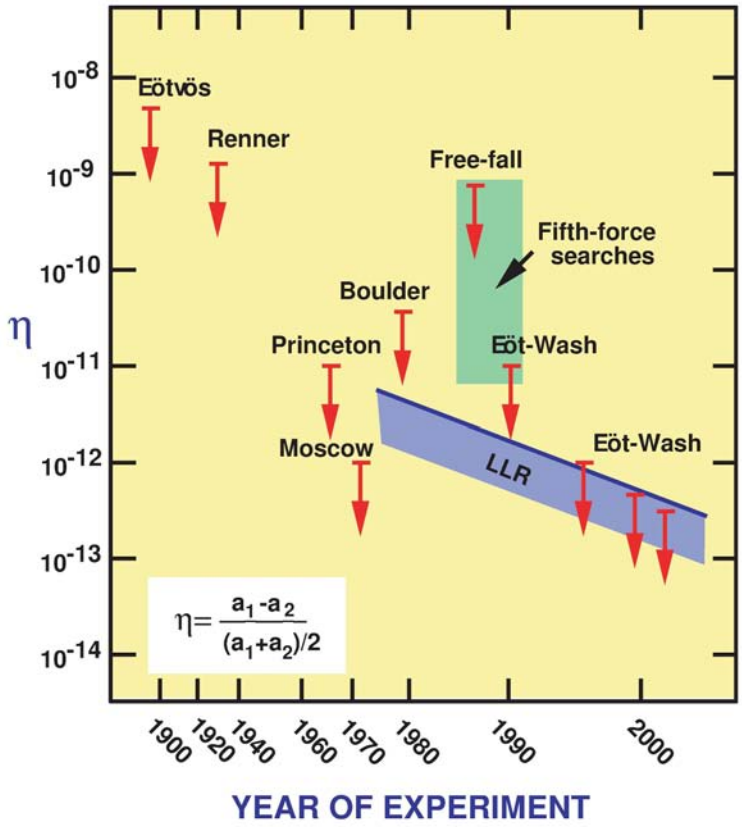

FIGURE 2.1: Limits placed on the parameter η by different experimental tests of the Weak Equivalence Principle [SOURCE: Will, 2006]

the value of η to the order of 10^{-13}.

Dicke[9] used the term "weak equivalence principle" (WEP) to distinguish this interpretation of UFF—the strictly mechanical, Galilean equivalence principle—from what he called the "strong equivalence principle" (SEP) which posited total, not just mechanical equivalence in a gravitational field. After Dicke, Will defined the "Einstein equivalence principle" (EEP) which is the new formulation of the principle of equivalence that Einstein used in order to

University of Washington, and he mentions space projects that are currently (2016) being developed with the hope of pushing the limit to $\eta < 3 \times 10^{-15}$.

[9]See for example *The roots of scalar-tensor theory: an approximate history*, p. 9, where we are told by Brans that, "Dicke often pointed out that we need to distinguish between two forms" which Dicke called WEP and SEP.

develop GTR. I examine exactly what EEP says and its consequences in the following subsection. However, I wish to make the distinction clear between the terms SEP and EEP as I use them in this work (in keeping with current common practice). As a more stringent requirement, SEP refers to the equivalence of *all* physical phenomena, including those involving self-gravitating systems such as stars and planets, and experiments involving gravity: Cavendish-type experiments.[10] In contrast EEP, which was a step Einstein found necessary in order to arrive at GTR, and is also part of all other metric theories of gravity, is less stringent and expressly excludes self-gravitating systems. Therefore, EEP refers only to the equivalence of all *non-gravitational* physical phenomena.

2.2.2 The Einstein Equivalence Principle

Although WEP is fundamental to GTR, it was not sufficient to allow Einstein to extend the principle of relativity from the special (inertial) case to the general one. In 1905, by basing STR on the constancy of the speed of light, he had already effectively adopted a new form of the principle of equivalence, although it did not become explicit until he went on to develop GTR. As John Stachel tells us when reviewing the process leading up to STR, instead of the strictly mechanical equivalence of Galileo and Newton, (and for the moment, excluding gravitational phenomena) Einstein required that:

> ... *all* the laws of physics take the same form in any inertial frame—in particular, the laws of electricity, magnetism, and optics in addition to those of mechanics.
>
> (Stachel, 2001, p. 160; italics in the original)

He was then able to use this new expression of equivalence to extend the principle of relativity to include accelerated motion and express all (non-gravitational) physical interactions in a form that was independent of the motion of the observer, whether inertial or not.

Einstein's first step from invariance under a uniform translation (the conditions of STR) to an invariant expression of physical theories under any

[10] As I mention in Subsection 2.4.3 below, it seems that adherence to SEP is a consequence of, and only applies to, GTR.

arbitrary motion was to consider the effect of uniform acceleration on massive bodies. He realised that rectilinear uniform acceleration—with its identical (mechanical) effects on all (massive) objects—could be equated, mechanically, to a homogenous gravitational field. His discovery in STR of the equivalence of mass and energy enabled him to take his next step and extend this idea (that uniform acceleration can be equated mechanically to a homogenous gravitational field) from the case of massive objects to include the effects on all forms of mass–energy; particularly, electromagnetic radiation.

In STR, physics is limited to *Lorentz* covariance; that is to say, only those relationships that are maintained under a Lorentz transformation can be considered physical laws: independent of the observer. Through the theory of invariables and the use of tensor calculus,[11] Einstein used *general* covariance to express physics in a form that would be valid no matter what transformations the system of reference underwent (not just the uniform translations described by Lorentz transformations). To be generally covariant means to be invariant under differentiable coordinate transformations; it ensures the validity of the mathematics of a theory when it undergoes a diffeomorphism[12] representing a transformation from one reference system to another in arbitrary motion relative to the first. It is important to realise that general covariance is a property of how a theory is expressed rather than the actual content of that theory. Einstein presented his final theory as generally covariant and stressed the importance of this characteristic. This has often been seen as the most important feature of the theory; especially in the years soon after the appearance of GTR. However, the importance of this condition has come to be questioned. Indeed, since Einstein's theory was published, it has been shown that Newtonian gravity (and possibly any spacetime theory) can be expressed in a generally covariant form. Norton tells us how, at the start of the 21st century, this is the new accepted view:

> A dissident view, however, tracing back at least to objections raised by Erich Kretschmann in 1917, holds that there is no physical content in Einstein's demand for general covariance. That

[11] At the time called "absolute differential calculus."

[12] A diffeomorphism is defined as a one-to-one and onto map of class C^∞ whose inverse is also of class C^∞.

> dissident view has grown into the mainstream. Many accounts of general relativity no longer even mention a principle or requirement of general covariance.
>
> (Norton, 2010, p. 110)

Despite this new mainstream view, Norton claims that the general covariance of GTR *is* physically significant, although it is not of central importance. He suggests that it is only through careful manipulation involving subtle changes in our interpretation of the physical reality that we manage to express Newtonian mechanics in a covariant form. Conversely, the property of the relations between physical phenomena that Einstein was searching for the means to express—that is, their independence from any fixed, absolute background—is precisely what is expressed in the formal general covariance of the theory. The important point about GTR is that it is most simply expressed in a covariant form; and whereas we may express other theories in a covariant form, it may well make them far more complicated. The simplicity of the generally covariant formulation of GTR demonstrates that this feature occupies a special place in the theory.

Apart from the requirement of general covariance, for the laws governing physical interactions to be independent of an arbitrary observer, they cannot depend on where or when an interaction takes place: the same experiment performed under the same conditions—but at a different time, or in a different place—must always yield the same results. If it is not only the metrical (gravitational) conditions that can affect the results, then we are not dealing with an independent physical law.

Einstein went on to extend the idea of the equivalence of gravity and acceleration from the particular (idealised) case of homogeneous gravity and uniform rectilinear acceleration to the status of a general principle. He combined it with his work on general covariance and was thus able to arrive at a theory of gravity as a geometrical condition of the spacetime around energy–momentum. It is clear from the time that elapsed between publishing STR, his first theory of gravitation in 1907, the idea of gravity as a geometrical phenomenon in 1912, and the finished GTR (1915) that it was a torturous task to develop these ideas in full. What we now know as EEP is embodied in his theory and can be seen as the total equivalence of gravity and acceleration

(and must, by definition, be shared by all metric theories of gravity). It thus includes (and explains) WEP. It furthermore includes the results of STR as just that; a (theoretical, idealised) particular model or instance in which gravity is absent; or equally, there is no acceleration.

Effectively, adopting EEP means that when the (non-gravitational) laws of physics are suitably expressed, they must take the form they have in STR. Furthermore, the results of all experiments must be the same wherever and whenever they are performed: there must be no preferences in spacetime location. These conditions are often known as local Lorentz invariance (LLI)—the reduction to STR in a laboratory that is in free fall—and local position invariance (LPI) or the universality of the gravitational redshift (UGR), which rather than implying that in the absence of acceleration all (non-gravitational) experiments yield the same predictions for an observer in any inertial frame, implies that spacetime is homogenous in all directions and therefore physical constants are constant everywhere and everywhen, so to speak.

It can be seen that GTR rests on, and builds on, the success of the idealised model we adopt in STR. The predictions of STR are regularly tested in many different ways in laboratory experiments. The accuracy of these tests can be considered as confirming LLI; one of the foundations on which GTR and all metric theories of gravity stand. As with the tests of WEP mentioned above, although these can in no way be considered as evidence for GTR over and above other metric theories, confirmation of STR does show that, in the limit as the local curved spacetime of GTR can be approximated to flat Minkowski spacetime, the theory is consistent with experimental evidence. Commenting on this aspect of the degree of confidence afforded to STR and therefore transferred to metric theories of gravity, and after considering recent experimental advances, David Mattingly, writing in 2005, concluded that:

> ... over the last decade or two a tremendous amount of progress has been made in tests of Lorentz invariance. Currently, we have no experimental evidence that Lorentz symmetry is not an exact symmetry in nature. ... The question that must be asked at this juncture in regards to Lorentz invariance is: When have we tested enough?
>
> (Mattingly, 2005, p. 60)

Einstein always maintained that his principle of equivalence was central to GTR; however, it has proved to be a very controversial issue. Einstein used the famous *Gedankenexperiment* or thought experiment of a lift in space to show that acceleration could mimic the effects of gravity. However, as Einstein himself made perfectly clear, this example is limited to uniform rectilinear acceleration reproducing the effects of a theoretical homogenous gravitation field. (Strictly homogenous gravity would have to be source free, as all our observations show that gravity obeys an inverse square law and it is therefore impossible to produce a homogenous field except approximately or in an infinitesimal region.) Such a gravitational field is effectively some kind of an additional field laid over a flat background Minkowski spacetime (or an accelerated flat Minkowski spacetime). However, the definition of gravity contained within GTR is that of curved spacetime[13] and this definition is clearly in direct conflict with the possibility of there being a homogenous gravitational field superimposed on a flat Minkowski spacetime.

The important point is that this was a thought experiment devised by Einstein as a model; it represents an idealisation of the properties of actual gravity in the limit as gravity tends to zero or we consider volumes that tend to zero. The idea of an actual accelerated flat Minkowski spacetime clearly cannot represent a real example of gravitation, but it can be a useful model. However we choose to define actual gravitation, it is clear that it is not compatible with the homogeneous gravitational field of the lift thought experiment: in the real world it has no such manifestation. However, the thought experiment demonstrates how a model allowed Einstein to move forward and extend STR. Einstein possibly used this model to aid his own thought process, and certainly to suggest to others an intuitive understanding of the equivalence of gravity and acceleration.

[13]Gravity is defined in GTR in several different ways whose meaning is equivalent however it is expressed in each individual instance: tidal forces produced by the non-vanishing of the metric curvature (which does vanish in the absence of gravity leading to the flat Minkowski spacetime of STR), the fact that the Christoffel symbols (often referred to as "Christoffel symbols of the second kind" but also called "affine connections" and "connection coefficients") cannot be eliminated from the equations by a simple coordinate transformation, or the non-vanishing of the Riemann curvature tensor.

The apparent contradiction between the actual curved nature of gravity and the model Einstein used has led to confusion as to how EEP can be maintained in GTR. One way to save EEP is to claim that it only holds in the infinitesimal limit. The argument is that on the curved GTR manifold, STR holds at an (infinitesimal) point. This interpretation has been called the infinitesimal principle of equivalence. When faced with this reading of EEP, Einstein pointed out that at the level of the infinitely small, all continuous lines are straight and therefore it is impossible to distinguish geodesics from other straight lines. I see no contradiction between the actual manifest world and Einstein's thought experiment, so long as we use his model correctly and realise its intended use and limitations. The key to understanding Einstein's point is that when he evokes his lift and thereby EEP, he is not talking about arbitrary gravitational fields or even any actual example of gravity (with its inverse square law manifestation). He is very specifically referring to the precise, idealised, notion of homogeneous gravity. Although we can find no real example of total equivalence, this does not mean that the notion has no meaning. Einstein is adopting counterfactual reasoning to tell us that if there were a universe which contained homogenous gravity, then we would not be able to distinguish between uniform, rectilinear acceleration and that homogenous gravity.

This idea is analogous to thought experiments in electromagnetism which require us to imagine two infinite flat parallel charged plates which form between them a homogenous electrical field with no edge effects. However, in the real universe, in the same way as we talk of "points" and "instantaneous velocities" when we consider these to be mathematical extrapolations from actual empirical evidence, so we can use the idealised model of perfectly homogeneous gravity. At the same time we can maintain that actual gravity, associated with a specific mass-energy distribution, is, by its very nature, a curved phenomenon. To make a real (curved) arbitrary field disappear through a simple coordinate transformation would indeed require a reduction to the infinitesimal (in order to make the effects of curvature disappear) with Einstein's objection as noted above; although at any individual point we can define a flat Minkowski system of coordinates.

2.3 The Parametrised Post-Newtonian Formalism

Through considering possible deviations from strict equivalence of m_I and m_P for massive bodies (as opposed to test bodies in the laboratory), Nordtvedt, 1968[14] developed a formalism for comparing different metric theories of gravity. The (PPN) formalism, as it became known, was further developed by Will (in conjunction with Nordtvedt) who then published a full version of it in 1972. It is a phenomenological framework which allows different metric theories of gravity to be compared and tested through observation and experiment, within an approximation appropriate for all low-speed (compared to the speed of light) weak-field situations, such as those encountered in the Solar System. In the framework, all theories are assumed to adopt the same form; in this way they can be compared via the different values that each theory produces for a set of shared parameters. These parameters are generated through power series expansions of Newtonian physical identities with additional terms representing the "adjustments" required for calculations to agree with Solar System observations and experiments. They describe, approximately, the deviations from classical dynamics predicted by different theories of gravity.

The formalism assumes a spherically symmetric static source of gravitation with a strength characterised by:

$$m \equiv \frac{GM}{c^2} \qquad \text{(where } M \text{ is the source mass)} \tag{2.3}$$

and a general Riemannian exterior geometry given by

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu \tag{2.4}$$

[14] Will, 2006 says on page 25 that Nordtvedt was "extending earlier work by Eddington, Robertson and Schiff" and further tells us that: "The parameters γ and β are the usual Eddington-Robertson-Schiff parameters used to describe the 'classical' tests of GR, and are in some sense the most important; they are the only nonzero parameters in GR and scalar-tensor gravity."

The components of the spacetime metric can be expressed by the general power series:

$$\begin{aligned} g_{00} &= -1 + 2\alpha\left(\frac{m}{r}\right) - 2\beta\left(\frac{m}{r}\right)^2 + \dots \\ g_{0j} &= 0 \\ g_{jk} &= \left[1 + 2\gamma\left(\frac{m}{r}\right)\right]\delta_{jk} + \dots \end{aligned} \tag{2.5}$$

where α, β and γ are dimensionless constants of order 1, which depend on (and are defined by) each individual theory of gravity; and r is a radial variable ($r^2 = x^2 + y^2 + z^2$). In the limit as r tends to infinity, the equations giving the metric are required to yield the flat, Minkowski (Lorentz) metric; as in the case of equations 2.5. With the given relationship for m and the source rest mass M, in order to give the Newtonian values in the limit as the field tends to zero, α must be equal to 1. Since Newtonian gravity describes Solar System phenomena very well and only deviates from observation as we move away from the weak-field approximation, it is totally reasonable to insist that all admissible theories must satisfy this condition, and α is therefore set equal to 1 and omitted from all the equations.

In order to compare field strengths in different situations, a characteristic parameter for the field strength, the compactness:

$$\frac{E_{grav}}{M_\circ c^2} = \frac{GM_\circ}{c^2 R_\circ}$$

is used. At the surface of the Sun, this parameter has a value of the order of 10^{-6} compared to a value of approximately 0.2 at the surface of a neutron star or 0.5 at the event horizon of a black hole. It could be the case that all the PPN parameters could be expressed as containing weak and strong field contributions, of the form $\alpha_{TOT} = \alpha_{WEAK} + \alpha_{STRONG}(c_1, c_2, c_3, \dots) + \dots$ where the factors $c_1, c_2, c_3, \dots$ on which the strong-field contribution depends, in turn depend on the compactness of the gravitational source, and that thus they could be used to accurately model gravitating systems that are beyond the usual weak-field approximation. It is clear, however, that contributions to the total parameter from the strong field need not be of concern when working in

the Solar System; or at least, not until we come to require an accuracy for the parameters of greater than 1 part in 10^6.

Within the framework, the parameter γ is interpreted as representing the amount of curvature produced in space by the presence of the source mass M, at a radius r. The parameter β is the amount of non-linearity that the theory predicts for the g_{00} component of the metric. In the full PPN used to compare scalar-tensor and other theories of gravity, there are 10 parameters in total (once the original α is disregarded as explained above), which are denoted: γ and β, (which have the interpretations I explain above); then ξ (which represents preferred location effects); α_1, α_2 and α_3 (preferred frame effects); and ζ_1, ζ_2, ζ_3 and ζ_4 (violation of conservation of momentum).[15] By definition, GTR is the theory for which $\gamma = \beta = 1$ and $\xi = \alpha_1 = \alpha_2 = \alpha_3 = \zeta_1 = \zeta_2 = \zeta_3 = \zeta_4 = 0$. Since only γ and β appear in the equations of motion for the gyroscopes that GP-B was designed to test, and furthermore the other 8 parameters take values of zero in GTR, I do not consider them further in this work.[16]

Although the PPN has proved to be very useful, it is important to be aware of its limitations. One obvious limitation is that assuming the field to be static and approximating it by a simple series expansion can tell us little about a system of rapidly changing fields. Certain theories (such as Rosen's bimetric theory) give results that are very similar to GTR within the limits of the PPN framework, but which in fact vary greatly under different assumptions. Thus, the qualitative value of the limited Solar System tests can be questioned, and may be compared to:

> ... studying the behaviour of a function, say $f(x)$, in a small neighbourhood of one point, say $x = 0$. Seen from this point of view, the general PPN expansion is analogous to parametrising, near $x = 0$, the behaviour of a general class of functions by means, say, of a parabolic approximation, $f(x) = \alpha + \beta x + \gamma x^2 + O(x^3)$. Clearly such a local parametrisation of $f(x)$ is unable to distinguish

[15] In fact α_3 represents both a violation of the conservation of momentum and preferred frame effects.

[16] They do appear in a quote from Will in Subsection 2.4.1, and I mention them again briefly in Subsection 2.6.2 when I discuss the Nordtvedt effect.

> among functions which approximate each other closely at $x = 0$, but behave very differently in the large.
>
> (Damour, 1992, p. S56)

While this criticism is certainly perfectly valid and should be borne in mind, the static, weak-field approximation used in the PPN framework is capable of distinguishing between the predictions of many different metric theories. Although this is only a first approximation and it will become necessary to go beyond it, it has proved very useful and has allowed many different tests of theories of gravitation to be developed with a minimum of theoretical assumptions.

This is directly related to our conception and use of models to represent the external world. If we select and model only that part of the phenomena that we believe affect our local surroundings directly, we can be tempted to affirm that we have devised a rigorous test for a theory, while in fact certain effects have simply not been included in the original set-up of our models. It is easy to consider that since the PPN has been devised and used for the low-speed, weak-field situation found in the Solar System, this environment provides an exhaustive testing ground for the parameters involved. This, however, may prove to be an illusion brought about by the idealisation that went into the design of the model.

2.4 Einstein's Three "Classic"[17] Tests of GTR

When Einstein published his review of GTR in 1916, he cited 3 observable consequences that he believed resulted directly from the theory. One of these, the precession of the perihelion of Mercury, had already been observed and as Einstein noted, his theory accounted for it perfectly.[18] This is the only one of the three that deals with relativistic effects on massive bodies; the other two

[17] I use "classic" here in the traditional way to describe Einstein's 3 tests which Schiff, 1960 (p. 340) calls "the three 'crucial tests'". I do not follow Will, 1993a in omitting the gravitational redshift from this classification and including instead the Shapiro time delay (which I consider in Subsection 2.6.1).

[18] See Einstein, 1952[1916], p. 200.

(the deflection of light by a massive body and the gravitational redshift) affect light. For this reason, it has been claimed that only the effects on Mercury can be considered to be true tests of GTR; with the other two testing only the weaker constraints of EEP. However, it is now generally accepted that while the redshift does indeed only test EEP, the deflection of light tests Einstein's theory in full. In this section, I consider the three tests, starting with the case of Mercury, then in 2.4.2 looking briefly at light deflection and concluding with the gravitational redshift in 2.4.3.

2.4.1 The anomalous advance of the perihelion of Mercury

Measurement of the rate of precession of the perihelion of Mercury could be regarded as the first empirical evidence in favour of GTR. However, as the effect was known and had been measured accurately before the theory was published, it can be argued that GTR did not predict it, but was designed to account for it. This argument tries to capture the common-sense interpretation of novel evidence as bearing more weight and thus supporting a theory more (or in this version of the argument, offering support instead of none) than evidence, such as Mercury's anomalous motion, that was known before the theory took shape. This is precisely an argument that I will take up, engage in and offer a certain resolution to, in Chapter 5. Therefore, I will not go into all the details of it here, but I will say that current interpretations of "novel," when referring to evidence in such circumstances, are interpreted as meaning novel with respect to the theory or the process of developing the theory. On one hand, there can be little doubt that Einstein's motivation for and his working towards GTR was not driven by a desire to resolve the mystery of Mercury's motion. On the other hand, however, we do not know if Einstein would have presented his theory when he did if the results it yields had not matched perfectly the observed anomaly. However, given the confidence he expressed in other, as yet untested, results of his theory also matching observation perfectly, I think there can be little doubt that he had certainly not designed a theory to match prior observations; though, of course, we do not know how his confidence may have been shaken had it not. The other facet of the novelty argument is that in terms of epistemic warrant there really is no

difference between the timing of the observations; the fact that a theory yields results that match observation is always evidence in favour of that theory (hence the recourse to the "common sense" value of novelty).

Many theories had been proposed that accounted for the observed anomaly, though almost all of them disagreed with other observations and were complicated additions to Newtonian theory which attempted to make an exception of this one observation.[19] This is certainly not the case for GTR, as the calculation of the correct (observed) perihelion shift for Mercury is a direct consequence of the theory, which makes no special case for it and requires no awkward additions to explain it. This result has not always been accepted as evidence in favour of GTR though. An alternative cause could be a solar gravitational multipole moment. Until solar observations and our understanding of the physics of the Sun were refined enough to allow this possibility to be ruled out, the possibility of such a perturbing effect (which just happened to coincide with the prediction of GTR) was maintained by some as a reason to disqualify this measurement as a test of GTR. Today, both theory and measurement have led to the conclusion that if the Sun has a quadrupole moment, it is too small to account for the observed anomaly.

Although Mercury is the closest planet to the Sun, it can still be described using a weak-field, low-speed approximation. Like all Solar System observations, the motion of Mercury is described well by Newtonian mechanics and the small (43 seconds of arc per century) deviation from the predictions of classical Keplerian motion requires only slight relativistic adjustments to Newtonian mechanics. However, the position of Mercury and the eccentricity of its orbit, mean that relativistic effects which are not appreciable in the behaviour of the other planets are important in this case. The PPN formalism is ideal for analysing the motion of the planet and within this framework we can say of the Mercury–Sun system that its:

[19] Proposed theories included: the existence of a new planet, called Vulcan, in an orbit close to the Sun; an additional ring of "planetoids" which would produce the perturbation; and a deviation from the inverse-square law of gravitation. Penrose, 2004 tells us that in 1894, Aspeth Hall proposed changing the power of 2 to 2.000 000 16 with quite some success!

> ... uniform center-of-mass motion is a property of fully conservative theories of gravity, whose parameters satisfy α_1, α_2, α_3, ζ_1, ζ_2, ζ_3 and $\zeta_4 \equiv 0$.
>
> (Will, 1993a, p. 111)

Even without this assumption of conservation, the ratio of the reduced mass of the system $\left(\frac{m_1 m_2}{m_1 + m_2}\right)$ to the solar mass is of the order of 10^{-7} and therefore these parameters, and also ξ, are negligible[20]. Only the degree of curvature of spacetime per unit rest mass and the deviation from linearity of the field equations affect Mercury's equations of motion.

We can then use the PPN in order to refine the Newtonian description of the acceleration of a body orbiting the Sun. In terms of the PPN parameters γ and β (which, as I explained in the previous section, can be seen as representing the degree of curvature of space in the presence of a mass source and the non-linearity of the theory, respectively; these being the two effects we need to include) we require, in addition to the Newtonian term $\left(\frac{GM_\circ \mathbf{r}}{r^3}\right)$, a "perturbative" term given in the PPN formulation by:

$$\delta a = \frac{GM_\circ}{r^3}\left[(2\gamma + \beta)\frac{GM_\circ \mathbf{r}}{r} - \gamma v^2 \mathbf{r} + 2(\gamma + 1)(\mathbf{r} \cdot \mathbf{v})\mathbf{v}\right] \quad [21] \tag{2.6}$$

It is clear that this can be divided into a component in the radial direction and a component perpendicular to this in the orbital plane, while the third component, normal to the orbital plane, is zero. By analysing the variation in the Keplerian orbital parameters, it is possible to arrive at the expression for the rate of change in the position of the perihelion per orbit (secular precession):

$$\dot{\omega} = \frac{3GM_\circ n}{a(1 - e^2)c^2}\left(\frac{2 + 2\gamma - \beta}{3}\right) \tag{2.7}$$

where a is the semi-major axis, e the eccentricity of the orbit and n is the mean motion. However, there is an additional term due to the effect of the (possible) solar quadrupole moment caused by the oblateness of the (rotating)

[20] See Balogh and Giampieri, 2002, p. 541.
[21] This is derived in Balogh and Giampieri, 2002, p. 541.

Sun. This effect adds an additional term to the above equation which depends on the Keplerian orbital parameters and also on the quadrupole moment J_2. The resulting quantity is: $\left(\frac{J_2}{10^{-7}}\right) 0.012''$ per century. The quadrupole moment of the Sun has not been measured directly, but current best estimates suggest that it is of the order of 10^{-7}, which is insignificant compared to the effect predicted by GTR, and in fact, below the expected limit of accuracy of the rate of precession. It can therefore be disregarded.

The empirical value of this precession is known to an accuracy of at least 0.1%.[22] This places a limit on the combination of parameters given by: $\left(\frac{2+2\gamma-\beta}{3}\right)$, which in 2006 was: $|2+2\gamma-\beta| < 3\times10^{-3}$ (from radar observations). It should be noted that observations of Mercury's orbit can only provide a limit for this combination of the two parameters. In order to calculate limits on each separately it is necessary to observe different phenomena or perform experiments.

2.4.2 Deflection of light by a massive body

In his final theory, Einstein gave a value of 1.7" for the bending of a ray of light grazing the Sun; another of the consequences his theory predicted. This value was a correction to the one he had published in 1911. On that previous occasion, he had based his calculation on a flat background spacetime (or relative space) and so, although he had arrived at the correct value compared to (asymptotic, distant) locally straight lines, his calculation failed to take account of the warping of spacetime near the massive body. This effect is responsible for half of the observed effect, and thus his correct value given by the final GTR was twice the previous value he had given.

[22] Will, 2006 gives this limit for data collected between 1966 and 1990, but he states that analysis of data from after 1990 could refine the accuracy further. In 2014, he opts for adopting $\gamma = 1$, in agreement with the latest data obtained from analysis of the Cassini spacecraft which gives this as accurate to within 0.001%, and then giving $1-\beta = (4.1 \pm 7.8)\times10^{-5}$ or $1-\beta = (0.4\pm2.4)\times10^{-4}$, depending on the source of the data.

Nowhere in his calculations does geodetic motion nor the full field equations make an express appearance. This has led some to question whether this effect can indeed be considered to be a test of the full GTR or whether (in common with the gravitational redshift, as in Subsection 2.4.3 below) it should be seen as a less stringent criterion testing only EEP. This is a deceptive argument since the equations of motion in GTR are not independent postulates; on the contrary, they are derived directly from the field equations. So despite the fact that we make no express use of the full theory, this does not change the fact that this perturbation arises from the field equations of GTR; and in this sense does indeed require the whole broader theory and its more in-depth aspects. In the PPN model, the factor that was missing in 1911 is supplied by adding in half the value of γ, which represents the curvature of spacetime and is different for each theory of gravity. The condition that $\gamma = 1$ goes beyond the requirements of EEP. The use of the PPN framework therefore shows that this is indeed a test of GTR.

It was Eddington's spectacularly publicised confirmation of this prediction in post-World War I 1919 that led to Einstein's meteoric rise to fame and the popular awareness of his GTR.[23] Although later results have led to the acceptance of the bending of light by massive bodies, it has been argued that Eddington's results were inconclusive and there were at the time no empirical grounds for such enthusiasm.[24] It is ironic that this observational result, which seems at the very least to have been overstated, was heralded as confirmation of the theory.

The techniques used for measuring this effect and the accuracy that can be achieved changed greatly with the advent of radio astronomy. Earth-orbiting optical telescopes have also improved on the results of traditional Earth-bound telescopes. The best estimates for γ in 2006 using deflection techniques came from using very-long-baseline interferometry (VLBI) and gave a value of $\gamma - 1 = -(1.7 \pm 4.5) \times 10^{-4}$. However, this result seems to have been improved

[23] For an account of how Eddington and his colleagues went about promoting GTR and the results of the 1919 expedition see Sponsell, 2002.

[24] In fact, in *The Golem*, Collins and Pinch, 1993 use this episode as an archetypal example of scientific discovery being led by social or institutional factors rather than by empirical evidence.

on recently by measurements of the Shapiro time delay using the Cassini satellite. The Shapiro time delay is a related effect which was unknown to Einstein and therefore I consider it separately in Section 2.6. Figure 2.2 shows the evolution since 1919 of the value and precision of γ resulting from different experiments using both photon deflection and time delay[25]

2.4.3 Gravitational redshift

In the final section of his 1916 paper, where Einstein talks of the observational consequences of his theory, he restates the first prediction he recognised as resulting from GTR.[26] In 1907 he had already predicted that the presence of a massive body would affect the metrical properties of spacetime causing a redshift in emission lines radiating from the body. In Einstein, 1952[1916], (section 22) he states that:

[25] It should be noted that as GP-B readied for launch, results were published that seemed to place tighter limits on the PPN parameters than GP-B aimed a. In response, the GP-B team has always defended the project by pointing out that whereas other observations may claim to provide tighter restrictions on these parameters, GP-B is an entirely different kettle of fish as it is a macroscopic experiment to test frame dragging and the geodetic effect. In GP-B we have a minutely observed system specifically designed precisely to measure these two effects predicted by GTR, as opposed to interpretations of distant natural phenomena or of the behaviour of photons within the Solar System. In his contribution to the 2015 Focus Issue of Classical and Quantum Gravity concentrating on spin, James Overduin (adopting what maybe considered the simplified lay term of the "source of the field") reminds us of the qualitative difference between GP-B and other tests of GTR:

> The importance of the geodetic and frame-dragging effects, however, does not lie in their implications for the PPN parameters. It lies in the fact that both these phenomena are qualitatively different from all preceding tests of GR, in that they depend on the spin of the test body and/or source of the field.
>
> (Overduin, 2015, p. 4)

[26] Will, 1993a comments on pages 67-8, that Einstein also predicted the bending of light by a massive body in the same 1907 paper, though he offered no calculation until one appeared (flawed) in his 1911 paper.

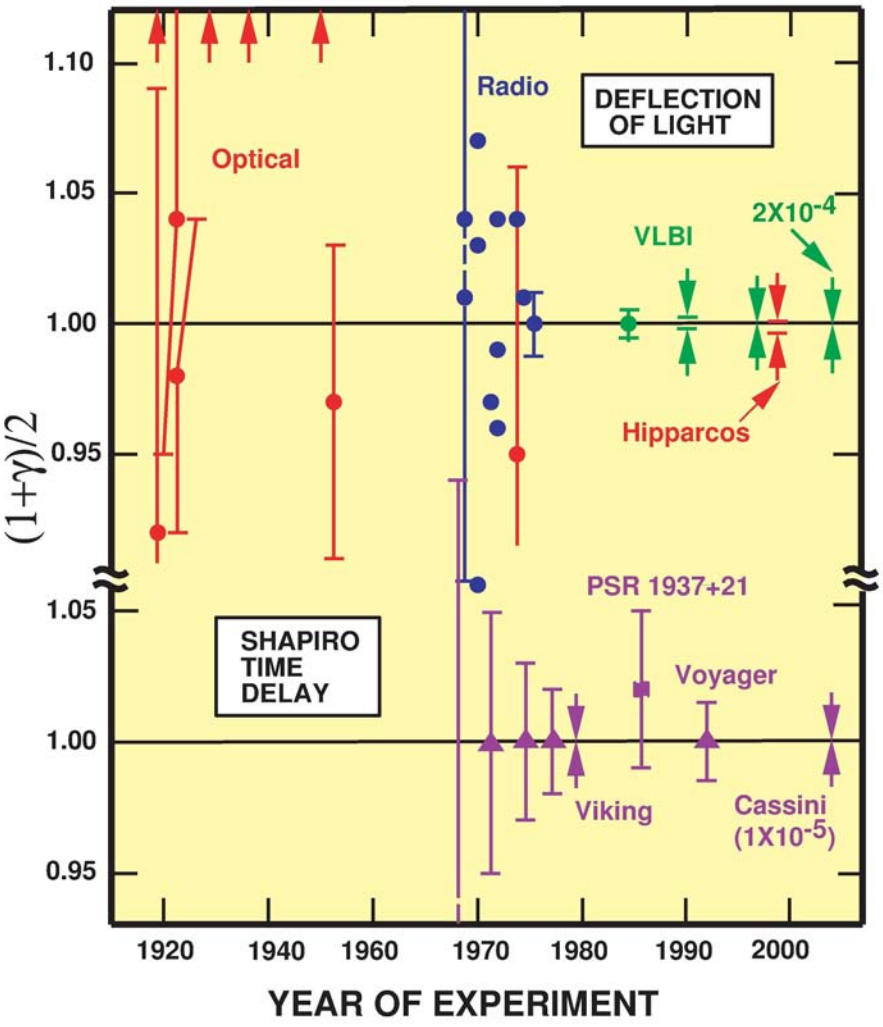

FIGURE 2.2: Estimates of the variation from $(\gamma + 1)/2 = 1$ based on different experimental results. Note that in the early years of the 21st century, right at the time of the GP-B space mission, both the VLBI result for the defection of light (2×10^{-4}) and the Cassini result for the Shapiro time delay (10^{-5}) gave results that placed a tighter limit on this parameter than GP-B aimed to achieve (10^{-3})
[SOURCE: Will, 2006]

> ... the clock goes more slowly if set up in the neighbourhood of ponderable masses. From this it follows that the spectral lines of light reaching us from the surface of large stars must appear displaced towards the red end of the spectrum.
>
> (Einstein, 1952[1916], p. 198)

Although Einstein considered this to be a test of GTR, this interpretation has been questioned and it is now accepted by most that the gravitational redshift cannot be considered as confirmation of the whole of GTR. Instead it is seen as evidence in favour of a broader aspect of GTR which it shares with other, later, theories: the metric nature of gravitation. Although Einstein clearly used the spacetime metric (or fundamental tensor, as he calls it in 1916) of GTR to calculate the redshift, there is no mention of geodetic motion or of the field equations that he derived earlier in the paper.

By definition, in all metric theories of gravity, the topology of spacetime is defined by a symmetric metric with a signature of 2 which is the representation of gravitation. Thus, energy–momentum and all non-gravitational fields respond only to this metric which accounts for all effects of gravitation (although there could be additional gravitational fields involved in determining it). Only metric theories are currently considered to be serious rivals of GTR as accurate formulations and representations of gravitation. This is due to the strength of the evidence in favour of STR, LLI and LPI for non-gravitating systems; and therefore, acceptance of EEP which is common to them all. The Brans-Dicke theory is the best-known rival theory of gravitation. Here the metric is determined by sources of energy–momentum (as in GTR) and also by an additional scalar field (which in turn depends on the metric, i.e., gravity). A different type of theory is Rosen's bimetric theory in which the physical metric is expressed as a field laid on top of, and depending on, a flat Minkowski background geometric metric. GTR is the only theory that contains no additional metric fields, and it is therefore the only theory for which the SEP holds (which, as I explain it in Subsection 2.2.1, states that EEP holds not only for non-gravitational effects, but also for gravitational—and therefore all physical—effects). This is due to the fact that if boundary conditions are chosen so that the metric, $g_{\mu\nu}$, becomes asymptotically flat at the limit of a region containing a gravitating system, we will not also be able to make the

additional fields disappear at the boundary. Therefore these conditions could influence the gravitating system via the values adopted by the auxiliary fields at the boundary. In this case the boundary would have to enter into our calculations and therefore, the behaviour of the system as a whole (including gravitational effects) would depend on its position within the universe, in contradiction to SEP. (Equivalence could still hold if we exclude the effects of gravitation on itself, as in the case of EEP.)

The fact that Einstein made the prediction of the gravitational redshift in 1907, armed only with EEP, would suggest that any theory which satisfies EEP (as all metric theories do) must share this prediction. In fact, it is easy to show (to the order of $\left(\frac{gh}{c^2}\right)$) using only STR and a (linear approximation to a) Newtonian gravitational potential ($\delta\phi(h) = \phi(R+h) - \phi(R) = gh$) that, at a small height above the Earth's surface (where h is the height above the Earth's surface and R the radius of the Earth), an absorbing atom placed above an emitting atom sees the frequency of the approaching photon as Doppler shifted (towards the red) by a frequency shift of: $\frac{\Delta v}{v} = -\frac{v}{c} = -\frac{gt}{c} = -\frac{gh}{c^2}$ where v is the velocity acquired by a free-falling observer initially at rest, in the time between emission and absorption. This effect cannot therefore be considered to test GTR.

It is quite easy to see that, since Einstein's theory was the only metric theory of gravitation at the time, he naturally considered that confirmation of the redshift he predicted would be valid evidence in favour of his theory. Adopting a Bayesian approach, such additional evidence would necessarily have to affect our degree of confidence in the theory; confirmation of the effect increasing the probability of the theory being correct (or strictly speaking, empirically adequate). However, with the advent of rival metric theories and the development of systems for comparing the predictions of different theories, it became clear that the gravitational redshift should be taken as evidence of the metric nature of gravity, rather than evidence for Einstein's GTR over and above other metric theories. In the Bayesian tradition, therefore, although confirmation would increase our confidence in GTR, it would likewise (seemingly paradoxically) increase our confidence in all rival theories, so long as they were metric theories. So although we can agree with Einstein that observation

of the predicted gravitational redshift is evidence in favour of GTR, this is so only insofar as it is a metric theory of gravitation; such evidence can do nothing to help us decide between rival metric theories of gravity.

What redshift experiments demonstrate is UGR. That is, that all clocks—periodic systems with an objectively definable frequency, such as vibrating atoms—regardless of their physical nature (and therefore their different degrees of dependence on various physical constants) demonstrate the same redshift, which is determined strictly and uniquely by the effective strength of the local gravity: the spacetime curvature; while at the same time of course, that curvature is determined by the presence of energy–momentum. There is a clear parallel to be drawn here with the manner in which tests of WEP (the universality of free fall) show that the mechanical effects of gravity are the same for all bodies, irrespective of their internal structure. Experiments demonstrating UGR show that gravity has an effect on all (non-gravitational) physical interactions (beyond the strictly mechanical effects of WEP) which is independent of where and when they are performed and therefore these experiments are also known as tests of LPI. As it is explained in a review of fundamental physics experiments planned for the international space station:

> The LPI principle of general relativity states that the outcome of any nongravitational experiment conducted in a local, free-falling frame is independent of where and when that experiment is conducted. A consequence is that different types of clocks keep exactly the same time, no matter where they are co-located in the universe, provided that the laws of physics do not vary from place to place.
>
> (Lammerzahl et al., 2004, p. 622)

In a similar way to the method of expressing possible departures from WEP explained in Subsection 2.2.1, we can define a parameter α such that the observed gravitational redshift, Z, is given by: $Z = (1 + \alpha)\frac{\Delta U}{c^2}$ where $\frac{\Delta U}{c^2}$ is the redshift predicted by GTR (and therefore all metric theories of gravitation). The parameter α therefore represents the observed degree of deviation from this prediction, due to the specific physical conditions of the experiment; that is, the physical processes involved in the clock—emission and absorption of the shifted signal—or preferred location effects. Results from gravitational

redshift experiments are therefore cited as placing limits on the magnitude of α.

In a famous experiment (almost legendary for students of relativity) involving the Jefferson Physical Laboratory tower at Harvard University, Rebka and Pound made the first experimental verification of the gravitational redshift. That was in 1960, just at the dawn of what Will has termed the "experimental age" of GTR[27] and coincided with the publication of Schiff's proposal for a gyroscope experiment which was to become GP-B. In that original 1960 experiment, Rebka and Pound achieved an accuracy of about 10%, which was improved on by a factor of ten over the following five years.

The most accurate measurement to date of the redshift was made in 1976 when a hydrogen maser atomic clock was launched in a rocket and flew for just under 2 hours in an elliptical path over the Atlantic Ocean. This experiment, in which NASA collaborated, was called Gravity Probe A. During the experiment, the change in the gravitational potential with the changing height of the clock was calculated and the clock signal was compared to that of an identical clock on the ground. The gravitational redshift was separated out from Doppler and STR effects. The accuracy of the experimental setup and the value obtained for the redshift set a limit of $|\alpha| \leq 2 \times 10^{-4}$. Since Gravity Probe A, high precision "null result" experiments have been performed on different types of atomic clock to show that they agree. This demonstrates that there is no non-gravitational redshift (that is to say that their composition and therefore dependence on different physical processes does not affect their rate, which is determined only by the gravitational potential in which they are situated). Current best results give a limit of $|\alpha_1 - \alpha_2| \leq 2.1 \times 10^{-5}$, where the difference $|\alpha_1 - \alpha_2|$ is between the two different clocks. Figure 2.3 shows the evolution of the limit on the magnitude of α resulting from selected experiments.

Although techniques for measuring the gravitational redshift from solar spectra have improved greatly, Einstein's original idea of using this technique has

[27] Will, 1993b distinguishes three stages in the history of GTR, with the experimental era commencing at the start of the 1960s due to a combination of theoretical and technological advances.

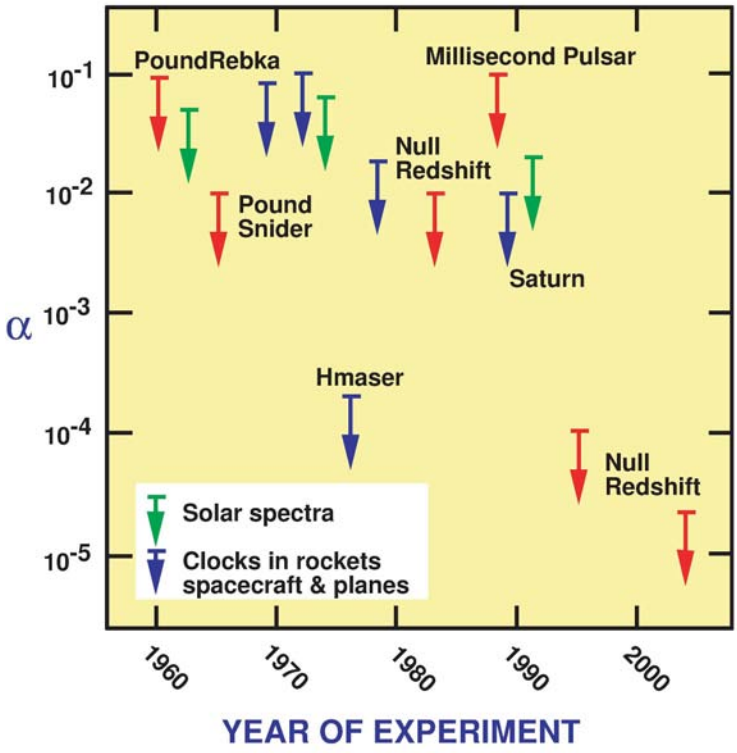

FIGURE 2.3: Estimates of the limit on α resulting from tests of the Universality of Gravitational Redshift (or LPI).
[SOURCE: Will, 2006]

only been able to yield accuracies of the order of 1%, due to the low signal-to-noise ratio resulting from the many complicated physical processes involved in their production and emission. Experiments and analysis of historical data have also been used to investigate whether fundamental physical constants vary over time; such variation could also indicate a violation of EEP as the redshift may not be universal with respect to time. There is a very wide range of experiments and data analysis, and the results are only tentative and very closely related to the complex theoretical models used. However, in the future this line of investigation may provide evidence for or against extra dimensions and it will possibly constitute the first empirically significant prediction of string theory. At the moment results seem to suggest a limit for the rate of variation of fundamental constants of the order of 10^{-15} or 10^{-16} per year,

though they are extremely tentative. [28]

2.5 The (de Sitter) Geodetic Effect and (Lense-Thirring) Frame Dragging

The calculation of the predicted geodetic effect was first published in 1916 by de Sitter[29] and the predicted frame dragging calculation was published shortly afterwards in 1918 by Lense and Thirring. Both of these effects should cause the precession of a gyroscope under the influence of gravitation. Both effects represent very small deviations from Newtonian mechanics which cannot normally be detected on the surface of the Earth. Therefore, despite being predicted by GTR and the calculations of their effects being published so soon, there was little hope of actually measuring them until extraterrestrial missions had been launched, artificial satellites had been placed in orbit and the prospect of free-fall experiments orbiting the Earth was a reality. It was thus in 1960 (shortly after the first artificial satellite, Sputnik, was successfully launched and its terrestrial orbit tracked in October 1957) that Leonard Schiff published an article presenting details of a "new" test for GTR based on these effects (George Pugh independently arrived at the same possibility at roughly the same time). [30]

[28] For a detailed review of the theoretical models used in this work and the results and conclusions reached see Uzan, 2003.

[29] He predicted an effect on the Earth–Moon system as it orbits the Sun. According to Keiser, 2003, p. 1212, the first confirmation of the effect, in agreement with GTR, was published in 1987 using lunar laser ranging data collected over a period of nearly 19 years. The accuracy of the measured effect was 2%, and this accuracy was later improved to 0.1%.

[30] The 1959 entry of the Project Timeline from the Stanford GP-B website tells us that: "Pugh's paper, 'Proposal for a Satellite test of the Coriolis Prediction of General Relativity', 1959, Pentagon Weapons System Evaluation Group (WSEG) memo #11, was never published in any public source until the year 2003." However, "Schiff's calculations were published in 1960 in his paper 'Motion of a Gyroscope according to Einstein's theory of Gravitation', L. I. Schiff, from the Proc. Nat. Acad. Sci. 46, pp. 871-882 (1960); also Phys. Rev. Lett. 4, pp. 215-219

There are 2 possible "perturbing" interactions between a spinning body, such as a gyroscope, and its motion under the effects of gravity: the effect of the body's spin on its motion through spacetime; and reciprocally, the effect of the body's motion through spacetime on its spin. Will, 1993a assures us, on page 163, that for a gyroscope with a radius of 4 cm (at about 200 rpm) orbiting the Earth, as in GP-B, deviations caused by the former consideration are at most of the order of 10^{-20} times the body's Newtonian acceleration; and can therefore be ignored. However, this is not the case for the latter effect of the gyroscope's orbital motion on its spin axis.

To calculate the precession of a gyroscope, Fermi-Walker transport is assumed. We can define the Fermi-Walker transport of a 4-vector as:

$$\nabla_{\mathbf{u}}\mathbf{S} = (\mathbf{S}\cdot\mathbf{a})\mathbf{u} - (\mathbf{S}\cdot\mathbf{u})\mathbf{a} \tag{2.8}$$

where $\mathbf{S}$ is the 4-vector representing the intrinsic angular momentum, or spin, of the body, $\mathbf{u}$ is its 4-velocity—that is, a vector tangent to its world line and of unit length—and $\mathbf{a}$ is the 4-acceleration of the body. Using this assumption produces totally consistent calculations and it is justified by its definition.[31] In the case of a gyroscope with spin as defined above (making the appropriate post-Newtonian, weak-field low-speed assumptions), the result of the Fermi-Walker transport gives us:

$$u^{\nu}S^{\mu}{}_{;\nu} = u^{\mu}(a^{\nu}S_{\nu}) \tag{2.9}$$

where $a^{\mu} = u^{\nu}u^{\mu}{}_{;\mu}$ is the 4-acceleration of the gyroscope; since in the local, commoving Lorentz frame, the acceleration is zero by definition. In this local frame, the 0-component of the 4-spin vector $= u^{\mu}S_{\mu} = 0$, and we can therefore treat the spin vector as a purely spatial vector.

(1960)." The NASA funding report (1965) cites the project as having started at Stanford in 1961 http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/1966007789.pdf accessed on 11/06/06.

[31] Will, 1993a tells us that: "All calculations to date have shown that ... the spin is Fermi-Walker transported along its world line." (page 164). However, he fails to mention the experimental evidence for this assumption, and it does not seem clear how calculations without empirical evidence can show such a fact.

Since the components of the spin vector in the PPN coordinate frame are equal to those in the commoving frame to within second-order terms,[32] for our purposes (a low-speed, weak-field approximation) the spin can be treated as a purely spatial vector in either the frame commoving with the gyroscope or the PPN-coordinate frame. This 3-vector is given by the equations:

$$\frac{d\mathbf{S}}{d\tau} = \mathbf{\Omega} \otimes \mathbf{S} \tag{2.10}$$

$$\mathbf{\Omega} = -\frac{1}{2}\mathbf{v} \otimes \mathbf{a} - \frac{1}{2}\nabla \otimes \mathbf{g} + (\gamma + \frac{1}{2})\mathbf{v} \otimes \nabla U \tag{2.11}$$

$$\mathbf{g} = g_{0i}e_{\hat{i}} \tag{2.12}$$

where $\mathbf{S}$ is now the 3-vector representing the intrinsic angular momentum or spin, $\mathbf{a}$ is the spatial part of the 4-acceleration, which for a body in free fall is zero, and is its 3-velocity.

The approximate metric for the Earth is given by:

$$ds^2 = -(1 + 2U)dt^2 + (1 - 2U)\delta_{jk}dx^i dx^k - 4h_j dx^j dt \tag{2.13}$$

where U is the Newtonian gravitational potential given once again by $U = -\frac{Gm}{r}$, which is to the order of $\frac{v}{c}$, and the spatial vector $\mathbf{h}$ is expressed in terms of the angular momentum of the gravitating body and the position as $\mathbf{h} = \frac{\mathbf{J} \otimes \mathbf{r}}{r^3}$, which is to the order of $\left(\frac{v}{c}\right)^{\frac{3}{2}}$. Combining this with the term given in the PPN formalism[33] for g_{0i} gives us an expression of the form:

$$\mathbf{\Omega} = \mathbf{\Omega}_{THO} + \mathbf{\Omega}_{FD} + \mathbf{\Omega}_{Geo} + \ldots \tag{2.14}$$

The first term in this expression for $\mathbf{\Omega}$ represents the Thomas precession which is a purely special relativistic effect, dependent on the gyroscope's acceleration.

[32] See Will, 1993a, p. 165.
[33] See Table 4.1 in Will, 1993a.

Again, in the locally commoving frame of a gyroscope in free fall, there is no acceleration by definition, and this term vanishes. The second term represents the effects of frame dragging or "gravitomagnetic" precession (so called, as its form is similar to that of magnetism in electromagnetic interactions). In order to calculate the precession for a gyroscope in Earth orbit, the rate of change of the drift angle represented by this second term can be given[34] as:

$$\mathbf{\Omega}_{FD} = -\frac{1}{2}\left(1+\gamma+\frac{\alpha_1}{4}\right)\frac{G}{r^3c^2}(\mathbf{J}-3\hat{\mathbf{n}}(\hat{\mathbf{n}}\cdot\mathbf{J})) \tag{2.15}$$

where $\mathbf{J}$ is the Earth's angular momentum, $\hat{\mathbf{n}}$ is a unit vector in the orbital plane, so that the value of $(\hat{\mathbf{n}}\cdot\mathbf{J})$ depends on the specifics of the orbit. This then leads to the expression for the change in the direction of spin for a gyroscope in a circular polar orbit, with orbital period P and radius a, as:

$$\delta S = \frac{1}{4}\left(1+\gamma+\frac{\alpha_1}{4}\right)\left(\frac{P}{a^3}\mathbf{J}\otimes\mathbf{S}\right) \qquad \text{per orbit.} \tag{2.16}$$

For an angle between the spin vectors of the Earth and the gyroscope of ϕ, this gives an angular precession of:

$$\delta\theta \cong 0.05" \left(1+\gamma+\frac{\alpha_1}{4}\right)\left(\frac{R_\oplus}{a}\right)\sin\phi \qquad \text{arcseconds per year.} \tag{2.17}$$

The value for GP-B in a polar Earth orbit at an altitude of 640 (± 10) km is 0.039 arcseconds per year.

It is important to bear in mind that frame dragging is an entirely general relativistic effect with no corresponding effect or approximation in Newtonian mechanics. In this sense it is a qualitative test of the metric nature of gravitation; the mere existence of an effect indicates a departure from Newtonian mechanics. This effect is not limited to one specific model of GTR; the Kerr metric is the framework that is usually used to calculate the effects, but it

[34] See Will, 2006.

appears to the same degree in all solutions or models that deal with rotating systems.[35]

The third term in the expression for $\mathbf{\Omega}$ above represents the geodetic effect (or, by analogy, the "gravitoelectric" effect). In order to calculate the precession for a gyroscope in an Earth orbit, it can be given[36] as:

$$\mathbf{\Omega}_{Geo} = -\frac{1}{2}(1+2\gamma)\mathbf{v} \otimes \frac{Gm\mathbf{r}}{r^3c^2} \tag{2.18}$$

This then leads to the expression for the change in the direction of spin for a gyroscope in a circular orbit of:

$$\delta\mathbf{S} = -2\pi\left(\gamma + \frac{1}{2}\right)\left(\frac{m_\oplus}{a}\right)\mathbf{S} \otimes \hat{\mathbf{h}} \qquad \text{per orbit} \tag{2.19}$$

which leads to an angular precession of:

$$\delta\theta \cong 8.1\text{"}\left(\frac{2}{3}\gamma + \frac{1}{3}\right)\left(\frac{R_\oplus}{a}\right)^{\frac{5}{2}} \qquad \text{arcseconds per year.} \tag{2.20}$$

However this term requires an adjustment to allow for the oblateness of the Earth (approx. 0.01" per year) and there is an additional term due to the potential of the Sun (approx. 0.02" per year). The value for GP-B of the total geodetic effect is 6.6 arcseconds per year. This term is 2 orders of magnitude larger than the frame dragging effect and would clearly swamp it, if it were not for the fact that the two effects are at right angles to each other (see Figure 3.1 in the next chapter). By measuring the overall change in orientation of the spin axis, it is therefore possible to resolve it into the components of the two individual effects. It is the case, however, that due to the scale of the effect, the degree of accuracy to which the geodetic effect can be determined is far greater than that to which the frame dragging effect can be measured.

The remaining terms in equation 2.14 represent interactions between the gyroscope spin and other Earth and Sun effects, and are too small to have any significant effect on the experiment so can be ignored from the results.

[35] See Collas and Klein, 2004.
[36] See Will, 2006.

2.6 Further Tests

2.6.1 The Shapiro time delay

In 1964, Shapiro calculated that an electromagnetic signal making a round trip passing near a massive body would suffer a time delay. His discovery led to a new means for testing the PPN parameter γ. The situation is somewhat complex since there is no Newtonian equivalent and it is therefore impossible to detect a perturbation of the Newtonian value. However, it is possible to obtain results by comparing the difference in time delays between a signal passing close to the Sun (or some other sufficiently massive body) and a similar signal as the reflecting body moves away from conjunction. The effect is related to the deflection of light and it provides results for the same parameter. The current best results are $\gamma - 1 = 2.1 \pm 2.3 \times 10^{-5}$. This result is very significant to GP-B as it is of greater accuracy than the expected results. Some estimates of the variation from $\gamma = 1$ resulting from different experiments are shown in Figure 2.2.

2.6.2 The Nordtvedt effect

As I mention above in Section 2.3, Nordtvedt was jointly responsible for developing the PPN approach as a model which was suitable for investigating the effects of weak gravitation. As a result of his involvement in this project and the insight he gained through constructing and manipulating a model suitable for representing Solar System gravity, he discovered a new effect. This is a clear example of the opportunities that arise from model construction and manipulation that I mention in Chapter 1. The effect is fittingly named after him.

It was in fact at an early stage of the development of the PPN that Nordtvedt became aware that many metric theories actually predict a violation of the WEP. It would seem that the Eötvös-type experiments rule out a violation of WEP, but the effect that Nordtvedt discovered depends on a complicated combination of several of the PPN parameters and on the ratio of the gravitational

self-energy of a body to its mass. As the mass of a body increases, gravitational self-energy becomes proportionally more significant. The effect predicts: $\frac{m_P}{m_I} = 1 - \eta_N \left(\frac{E_g}{m}\right)$, where $\eta_N = 4\beta - \gamma - 3 - \frac{10}{3}\xi - \alpha_1 + \frac{2}{3}\alpha_2 - \frac{2}{3}\zeta_1 - \frac{1}{3}\zeta_2$. For laboratory objects, such as those used in Eötvös-type experiments, this value is less than 10^{-27} which is far too small an effect to be detected in the laboratory. However, for astronomical bodies the effect could be detectable. In fully conservative theories such as GTR all the PPN parameters are zero except for β and γ. If we limit consideration to these theories, then $\eta_N = 4\beta - \gamma - 3$. For GTR, in which β and γ are both equal to 1, $\eta_N = 0$.

However, if there were a deviation from GTR, and β and γ are not equal to 1, then the effect on the Earth would be greater than that on the Moon, due to the Earth's greater gravitational self-energy per unit mass. This would cause a perturbation of the Earth–Moon system as it orbits the Sun, resulting in a measurable effect on the Earth–Moon distance. Since 1969 when the first lunar reflector was put in place, measurements of the Earth–Moon distance have been made using lasers. The modelling that needs to be done in order to detect a change in the Earth–Moon distance is extremely complicated, but current best estimates give a limit of $|\eta_N| \leq 4.4 \pm 4.5 \times 10^{-4}$.

2.6.3 Strong gravity

For the sake of completeness I should mention that while strong gravity does not occur in the Solar System, that does not mean that we cannot observe its effects. The most important opportunities for this are binary pulsars, such as Hulse-Taylor—the first to be observed, in 1974. Different models are required for investigating such systems though clearly they afford a new opportunity to test the predictions of GTR.

The other important phenomenon is the possible detection of gravitational waves as they pass through the Solar System. This subject has always been highly controversial. As I mentioned at the beginning of this chapter, I hope to be able to continue research in this area in the future. For now, suffice it to quote Will on the possibilities for the not-too-distant future of the field:

> Some time in the next decade, a new opportunity for testing relativistic gravity will be realized, when a worldwide network of kilometre-scale, laser interferometric gravitational wave observatories in the U.S. (LIGO project), Europe (VIRGO and GEO600 projects) and Japan (TAMA300 project) begins regular detection and analysis of gravitational-wave signals from astrophysical sources. ... In addition to opening a new astronomical window, the detailed observation of gravitational waves by such observatories may provide the means to test general relativistic predictions for the polarization and speed of the waves, for gravitational radiation damping and for strong-field gravity.
>
> (Will, 2006, pp. 58-59)

Chapter 3

Gravity Probe B: An Experiment in General Relativity

"No mission could be simpler than GP-B; it's just a star,
a telescope and a spinning sphere"

Bill Fairbanks (GP-B Co-Founder)

3.1 Introduction: The Best Laid Plans ...

In this chapter, I consider the Gravity Probe B (GP-B) mission as it was conceived designed and executed: how it aimed to test GTR. GP-B could be seen as a typical experiment that just happened to have a very long lifespan, and was very well researched and documented. This alone would make it a useful case study as an example of experimental science. However, it was also an exceptional experiment in many other respects: an extremely rare exercise in fundamental aspects of experimental gravitation involving an extraordinary range of support and subsidiary work. For many people, the technical requirements alone were reason enough to justify completing the mission. Indeed, in the Focus Issue of Classical and Quantum Gravity dated 19th November 2015 that covers the history and results of GP-B in detail, Clifford Will in the Preface comments that fully 16 of the 21 papers contained in that issue "are primarily technological, describing everything from how the rotors were made and tested to the drag-free control of the spacecraft" (p. 4). However, I am not

overly concerned in this dissertation with the technical achievements of the mission; although I will return to the requirements for a successful mission as they were initially foreseen in Chapter 5, where I discuss the actual results in detail. Rather, in this chapter, I concentrate on the conception and aims of the scientific experiment itself; then later (in both Chapter 5 and Chapter 6), I consider how the actual results have been interpreted together with criticism of the (necessary) departure from the initial ideas I first explain here, and its possible justification.

In the introduction to the Gravity Probe B (GP-B) website[1] we are told that:

> ...although [Einstein's GTR] is among the most brilliant creations of the human mind, weaving together space, time, and gravitation, and bringing an understanding of such bizarre phenomena as black holes and the expanding Universe, it remains one of the least tested of scientific theories. General relativity is hard to reconcile with the rest of physics, and even within its own structure has weaknesses.
>
> (Everitt et al., 1993, quoted on the GP-B website) [2]

This immediately makes one wonder why it is considered such a masterpiece if it has inherent flaws, does not seem to fit in with the rest of physics and has hardly ever been tested! It certainly seems to justify attempts to test the theory further, and thereby provides a powerful argument in favour of GP-B. The critical Science Phase of the space mission during which the primary experimental data were collected was completed in 2005. The intention of the project team at the launch of the space mission in 2004 had been to then embark

[1] Throughout this chapter I make use of the GP-B website as a source of material. I accessed the website regularly during the mission itself to follow its progress and after the end of the mission, as problems were detected and work-arounds developed. It is not peer reviewed and clearly publicises the line of thought favoured by the project team. However, it does contain a considerable amount of general material concerning the mission, much of which has been prepared by leading experts. Furthermore, until the final results were released in 2011 and the Focus Issue of Classical and Quantum Gravity was released in 2015, apart from the more technical aspects of the project, it was quite difficult to find papers discussing the project in learned publications. The website has changed greatly over the years, and as I write in 2016, is still up and running, though much of the material I use here was accessed during the space mission and data processing period.

[2] See: http://einstein.stanford.edu/ accessed on 22/08/2006.

on a three-stage data analysis process, to last approximately a year, with final results to be announced around April 2007. During the three stages of that post-mission data analysis, the data already collected during the Science Phase were to be processed, results derived from them and interpreted, and conclusions drawn.

Most people who study GTR believe that some correction or adjustment will be needed to it; that is to say, it is not considered to be a "final theory" but rather our current best theory. In order to find out at what level and in what way any future adjustments to our fundamental physics theories will have to be made, we need to determine exactly how the manifest physical world deviates from our current best theoretical description of it: GTR, in the case of gravitation. It is therefore necessary to constantly improve and refine our knowledge, and one way to do this is to devise methods to provide more accurate experimental data; this was always the reasoning behind and the primary motivation for GP-B.[3]

As it turned out, the data collected were not at all as expected, and the team embarked on a long and drawn out process to try to establish why and how they had varied from expectations; and to develop the techniques necessary to salvage meaningful results from the exceptionally noisy dataset. At this stage, as I write in 2016, justifying the project does not seem too important or relevant; however, throughout much of its history, GP-B had to defend itself from those who would rather have seen the resources invested in it redirected to other areas. In 2008, already months after final results should have been published, after repeated delays and overspending, and over 40 years of involvement in the project, NASA pulled out of the project, abandoning the Stanford team to their fate. After temporarily surviving on private donations and an agreement for a bridging arrangement with NASA, the team found an alternative partner in the King Abdulaziz City of Science and Technology in Saudi Arabia. Nearly three years later, in May 2011, the final results were announced. Things had changed greatly since the space mission was launched on 20th April 2004. Much of the published material did not deal with the

[3] Although below, in Section 3.2, I note the political motivation for and military interest in the project.

test of GTR that had been the aim of the mission and which had inspired the entire project back in the early 1960s. Rather, the results that were published were also dedicated to explaining and justifying the techniques developed and the methods adopted to salvage meaningful results from the unexpected data actually recorded during the Science Phase of the space mission.

Similarly, my later analysis of the project will centre on the actual results of the mission, on this change in direction and on questions raised by the new techniques and data analysis process. However, in this chapter I consider the experiment as it was conceived, designed and executed, together with the expectations for it before and during the Science Phase; before the causes of the problems became apparent. A clear exposition of the experiment is necessary for us to be in a position to understand the problems that arose and above all, the solutions proposed and adopted, together with the criticism they provoked.

As I show in Chapter 2, the three tests proposed by Einstein do seem to have yielded results which confirm the empirical accuracy of his theory. So we may ask what is special about the frame-dragging and geodetic effects that GP-B set out to measure, and what their importance is when we were already told before the GP-B space mission that: "Over the past 90 years, various tests of the theory suggest that Einstein was on the right track"[4]. In fact, even the claims that have been made regarding the possible results of the GP-B experiment vary widely. In their paper (which appears to have been written originally in 1996), J. P. Turneaure et al. tell us, prior to launch, that:

> We expect GP-B 's measurement of the geodetic effect, which is related to the de Sitter effect, to improve the accuracy of non-null tests of general relativity by a factor of from 10 to 50. Also, the measurement of the frame-dragging effect will be the first direct measurement of this phenomenon, which is due to the dragging of space-time by the rotation of a massive body.
>
> (Turneaure et al., 2003, pp. 1387-8)

If no effects had been detected, would that really have cast a shadow of doubt on GTR when it was already hailed as one of the most important discoveries of the last century? We have to realise that the experiment formed part of

[4]From the Stanford University Gravity Probe B website.

an ongoing process of production and refinement of experimental results. As I mention above, it would be extremely naive to think that our current theories are final theories; rather they are provisional conclusions awaiting further refinement or even wholesale overthrow. Within this context of the inevitably transient nature of our (current) best theories, the experiment was set to provide useful and maybe crucial results.

GP-B was different from other attempts to verify GTR, in that it was a true physics experiment: not simply observations of systems that are beyond our control (whether totally, or only partially). It could therefore be adjusted and tuned and the hope was that unwanted effects could thus be eliminated or reduced to levels where they did not swamp the effects the experiment aimed to measure. However, as I aim to show in later chapters, it was precisely the extreme technical requirements that led to the overall results of the entire fifty-plus-year project being threatened. Again, this may cause us to question whether the mere fact of being an experiment could ever make it more reliable than the "simple" observing of the heavenly bodies carried out, for example, in the verifications of the bending of light near a massive body and the precession of the perihelion of Mercury. It could be the case that precisely the required degree of design and control is more likely to lead to results that are less objective, less rigorous and more a product of our design and interpretation. The data that GP-B collected was far removed from the final results and conclusions that were drawn from them; and epistemic dangers such as that of over-fitting the data to the desired model or allowing experimental design to dictate results must be taken seriously if the experiment and the conclusions drawn from it are to stand up to critical examination.

In Section 3.2, I review some of the history of the gyroscope experiment, as GP-B was originally known, and try to place the decision to go ahead with the experiment in a broad social and historical context. In Section 3.3, I describe the specific aims of the experiment itself: the two effects it aimed to quantify. In that section I also discuss some of the implications of the original experimental method, and consider how the technical solutions that were adopted—both in the physical set-up and in the data processing—affect the credibility of the results. Then, in Section 3.4, I explain in some detail exactly

what GP-B attempted to measure in order to achieve its overall aim. Finally, Section 3.5 is dedicated to a discussion of the possible results as they were envisaged prior to the execution of the mission, and the potential conclusions they may have led to. It is not until Chapter 5 that I go into greater detail of the specific experimental requirements for a successful experiment; there I will also present and analyse the actual results and contrast them to the expectations I explain in the present chapter. It is vital to understand the deviation of the results from initial expectations in order to really come to terms with the way in which the post-space mission data analysis unfolded and the departure the process represented from what had been foreseen prior to the launch of the rocket that placed the GP-B satellite in its near-Earth orbit on 20th April 2014.

3.2 History and Background of Gravity Probe B

The main elements of the original Stanford Gyroscope Experiment[5] that was to become GP-B appear to be quite straightforward and did not change significantly after its inception at the end of the 1950s. They are: a gyroscope inside a satellite in a purely gravitational orbit around the Earth; a telescope attached to the satellite enclosing the gyroscope and aligned with a suitable guide star; and finally, a method of comparing the direction of the telescope and the direction of the spin axis of the (free-falling) gyroscope.[6] A general

[5]When listing William (Bill) Fairbank's contributions to the advance of experimentation, Everitt in 1988 refers to this first of his projects after moving to Stanford in 1959 as the "NASA gyroscope experiment" (p. 22). In contrast, in Chapter VI of the same volume, Kip S. Thorne titles his contribution "Gravitomagnetism, Jets in Quasars, and the Stanford Gyroscope Experiment" (p. 573) while each of the 6 subsequent sections of that chapter refers in its title to some aspect of "the Stanford Relativity Gyroscope Experiment". So, it is clear that many variations on the name were used before GP-B was finally adopted.

[6]Let me just note that through the use of the term "enclosing" in the second of these elements, and despite having said in the first that the gyroscope is "inside" a satellite, I wish to emphasise the point that the gyroscope is in free fall and therefore although the satellite surrounds the free-falling gyroscope rotor, in extremely close proximity to it, and indeed shares its orbital trajectory, the satellite does not "contain" the gyro in the sense of restricting it or limiting its movement in any way.

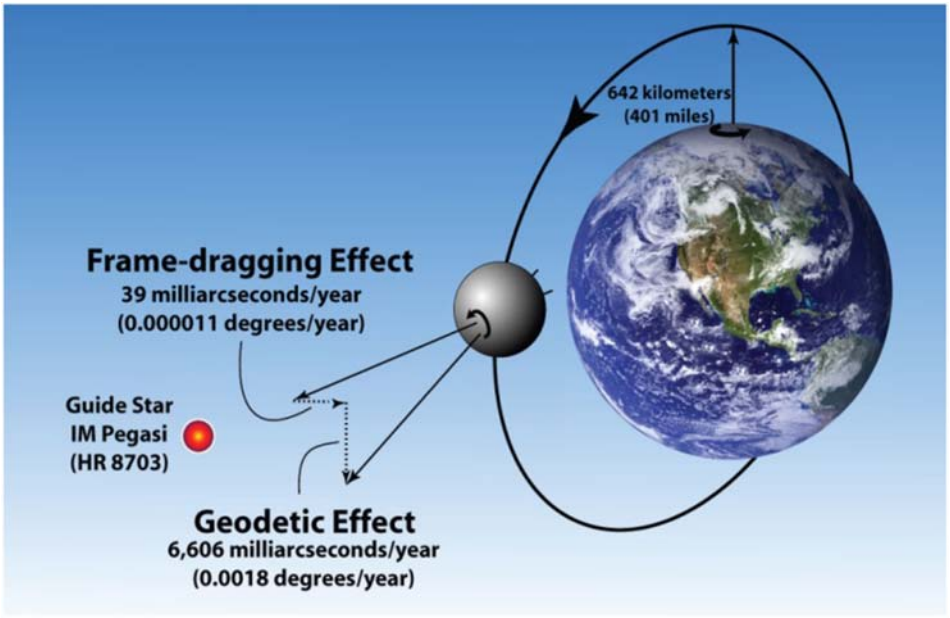

FIGURE 3.1: Representation of a GP-B gyroscope in a polar near-Earth orbit, showing the initial direction of the gyro spin axis pointing towards the guide star and the shift in that orientation predicted by GTR for the specific set-up.
[SOURCE: Everitt et al., 2015]

idea of the set-up is shown in Figure 3.1, but just illustrating a gyroscope and the predicted shift in its spin axis orientation without any of the experimental hardware. Despite this apparent simplicity, the difficulties that had to be overcome before the mission could finally be launched on 20th April 2004 were, to a large extent, due to the degree of precision required of the experiment if it was to be a rigorous test of GTR. While this was indeed the case for many of the specific claims made concerning the experiment, in contrast some of the more ambitious claims were based in a purely qualitative way on the theory behind the experiment. Thus, we see that while some made the very modest claim that GP-B would increase by a factor of ten the level of accuracy to which we knew the factor in the PPN framework, others claimed that it was the first true (or even the ultimate) test of GTR.

As I mention in the previous chapter, a space gyroscope experiment was first

suggested around 1960 by L. I. Schiff and independently by G. E. Pugh.[7] However, as far back as the 1930s, P. M. S. Blackett, who was later to supervise the research carried out by C. W. F. (Francis) Everitt—the GP-B Principal Investigator—for his PhD, had considered the idea of building an Earth-bound gyroscope to measure the frame-dragging effect, but had decided that it was not feasible.[8] In his 1960 paper, Schiff compared the behaviour of a gyroscope supported in a laboratory on Earth with that of a gyroscope in a free-falling satellite in a near-Earth orbit. He concluded that the observable frame-dragging and geodetic effects would be comparable per orbit for the two set-ups. This would mean that the orbiting gyroscope could gather data more quickly than the Earth-bound one; which would necessarily rotate just once every 24 hours. More importantly, however:

> ... most of the experimental difficulties that seem to rise [sic] with a high-precision gyroscope, especially instrumental torques, are greatly reduced if the gyroscope does not have to be supported against gravity.
>
> (Schiff, 1960, p. 882)

In the opening section of the introduction to his text, originally from 1959, Pugh comments that:

> ...the development of satellite technology will soon allow an experimental verification of the ...Lense-Thirring effect predicted by general relativity. ...Einstein remarked that the magnitude of this effect ..."is so small that confirmation by laboratory experiments is not to be thought of. However, artificial satellites can provide an almost force-free environment that cannot be approached in surface laboratories. This new environment makes the experimental confirmation of the prediction a distinct possibility.
>
> (Pugh, 2003[1959], pp. 415-416)

Despite the evident difficulties involved in placing a gyroscope in an artificial satellite and observing it in minute detail as it orbits the Earth, the idea behind a satellite-borne gyroscope experiment to be performed by Stanford University—where Schiff was head of the physics department—had been born. In his paper concerning the feasibility and requirements of the project,

[7] See footnote 30 in Chapter 2 above
[8] According to Fairbank et al., 1988, p. 588

Cannon (Cannon, 1963, p. 147) estimated that to meet the experimental requirements, the improvement required in the accuracy of the (then) best available gyroscope technology would have to increase drift performance (the stability of the gyroscope spin axis) by a factor of the order of 106. Years later, in the 1980s, Everitt would recall with affection the talk in 1961 of a "preposterously low drift rate of 1/100 of an arc-second per year" (Fairbank et al., 1988, p. 20). As both Schiff and Pugh had realised, the most feasible way to achieve this was the immensely difficult task of monitoring a gyroscope orbiting the Earth in free fall. This option was more appealing for a variety of reasons. First, an orbiting gyroscope requires no forces to support it.[9] Furthermore, averaging over orbits allows many non-relativistic factors which affect Earth-bound gyroscopes to be averaged away. Commenting on the proposal in 1963, Cannon suggested that a ground-based installation could prove useful but would only stand a chance of obtaining meaningful results for the geodetic effect, which he refers to as the largest of the relevant motions:

> ... an earthbound experiment would provide valuable experience with a complete experimental system ... [and] ... it might be possible to measure the largest of the earth-spin-vector motions ...
> (Cannon, 1963, p. 155)

However, for a full experiment to be performed, it was clearly necessary to use a gyroscope in a satellite. In 1968, when he comments on the possibility of using a gyroscope experiment to compare GTR with the Brans-Dicke theory and considers Schiff's proposal, O'Connell—who had jointly published a paper on a closely related subject the previous year—comments simply and dryly that "the gyroscope in a satellite offers a more sensitive test than the Earth-bound gyroscope." (O'Connell, 1968, p. 70)

Given the enormity of the problems the mission faced, especially from the perspective of the 1960s, it may seem strange that this type of experiment was seriously considered at all. Potentially the same results were available from a more conventional laboratory experiment. However, the problems

[9] However, it may of course be the case that the actual practicalities of executing the experiment require some forces to be applied to the gyro to maintain it correctly positioned with respect to the instrumentation used to take readings; this will inevitably be the case if more than one gyro is used in a single spacecraft.

faced by an Earth-bound experiment would have been very different and no doubt, so too would the technologies required to overcome them. Although some of the people working in the field considered the orbiting experiment to be more practical, it seems to me that the decision to back exclusively the satellite experiment and invest only in that option was taken somewhat hastily; especially if, with the benefit of hindsight, we consider that the experiment was to take some 50 years to produce final results that even then hardly represented a major breakthrough in our understanding of the universe we live in.

One possible reason why the satellite experiment was so attractive can, I believe, be found precisely in this historical perspective. On 25th May 1961, in an address to the US Congress, US President John F. Kennedy famously upped the odds in the space race by committing the USA, "to achieving the goal, before this decade is out, of landing a man on the Moon and returning him safely to the Earth."[10] With the Cold War raging, and the USSR at that time clearly ahead in space exploration, the USA was understandably keen to seize any opportunity that came to hand to develop and demonstrate its technologies and know-how in space. It could also be argued that this is an example of "Big Science"[11] and expensive showcase projects being used not for their empirical, scientific value, but for other, politically motivated reasons. Initial acceptance and interest in the experiment was no doubt influenced by

[10] The speech is available on YouTube: https://www.youtube.com/watch?v=TUXuV7XbZvU accessed on 27/08/2016.

[11] Although this is not how I intend it here, "Big Science" is usually used as a derogatory term to conjure up an image of science as providing "jobs for the boys" being a "gravy train" and leading to lazy scientists and poor science. The term seems to have first been coined by Alvin Weinberg in 1961, when he typified it as consisting of "the huge rockets, the high-energy accelerators, the high-flux research reactors" which he characterised as "monuments" and "symbols of our time" (Weinberg, 1961, p. 161), which is how I use it here. This origin is supported by the study of the subject edited by Peter Galison (a colleague of Everitt's at Stanford) and Bruce Hevly, including a contribution by Everitt titled "Background to History: The Transition from Little Physics to Big Physics in the Gravity Probe B Relativity Gyroscope Program", published in 1992, where the authors of the final chapter tell us: "When Alvin Weinberg, almost thirty years ago, examined the phenomenon he labelled 'big science,' ... He warned that big science could attenuate science itself." (Kargon, Leslie, and Schoenberger, 1992, p. 335)

these factors. It should be noted that Schiff's 1960 paper contains a footnote to the title informing readers that it is "Supported in part by the U. S. Air Force";[12] while Pugh's text (written in 1959) was originally an internal memo of the Weapons Systems Evaluation Group of the Department of Defence at the Pentagon.

Once the initial idea had been researched further, it is clear that a very important turning point for the experiment came in 1964 when, as a: "Proposal to Develop a Zero-G, Drag-Free Satellite and to Perform a Gyro Test of General Relativity in a Satellite"[13] it acquired funding from NASA (which had been formed in 1958, chiefly—supposedly—to oversee civil space research). This can be seen as the end of a pre-production, "ideation" stage, and the beginning of the production stage of the experimental project in earnest. Such a shift from ideas to experimentation clearly required much more investment. The project received another funding boost twenty years later when, in 1984, the Lockheed Corporation (later to become the present-day Lockheed Martin) joined the team at Stanford University as the main subcontractor to build the experimental hardware. It is estimated that overall the project cost of the order of 1,000 million US dollars.[14]

Despite constant revision of the timetable for the execution of the project, the overall goal was never abandoned. On more than one occasion GP-B came close to being scrapped; but it always managed to survive. (A 2003 news article in the popular magazine Science News produced by the not-for-profit organisation Society for Science and the Public reported that, "NASA has cancelled and then reinstated the mission seven times".[15]) The sheer longevity of the project, together with the facts that it survived NASA budget cuts, it attracted private funding and it was regularly successfully defended before

[12] In Schiff, 1960, asterisked note to the title, appearing on page 882.

[13] https://einstein.stanford.edu/content/sci_papers/papers/GPB-NASA_Proposal-Nov1962.pdf accessed on 27/08/2016.

[14] This can be compared to the cost of an average commercial telecommunications satellite at about 400 million US dollars, or the total annual NASA budget which is currently in the region of 19,000 million US dollars (some 0.5% of the total USA federal budget) (https://en.wikipedia.org/wiki/Budget_of_NASA accessed on 25/08/2016).

[15] Peter Weis 2003 in "Science News" http://www.sciencenews.org/articles/20031101/bob9.asp] accessed on 25/08/2016

different panels of experts, strongly suggest that it was far more than an exercise in sabre rattling, one-upmanship or political propaganda.[16] Apart from the undeniable wealth of technological spin-offs that the project produced,[17] many experts and social commentators seem to have been convinced of the enormous value of the scientific project. For example, in 2011, when the final results of the project were eventually published, another article in Science News was titled "Gravity Probe B finally pays off".[18] To whatever extent initial enthusiasm was based on the prospect of developing new weapons systems and beating the USSR in at least one aspect of the space race, the launch of GP-B on 20th April 2004, more than a decade after the end of the Cold War and the collapse of the USA's only real rival in terms of militarism and space exploration, has to be regarded above all as a crucial step in the long history of a remarkable scientific experiment.

3.3 The Objective of the Experiment

Between publication of the original idea for a satellite-borne gyroscope experiment to test GTR in 1960[19] and the announcement of final results in 2011 (or maybe the appearance of the Focus Issue of Classical and Quantum Gravity dedicated to GP-B in 2015, the publication of which Clifford Will tells us on the very first page has the aim of "thus bringing to a close an extraordinary chapter in experimental gravitation" (Will, 2015, p. 1)) the entire project can

[16] It is worth noting, however, that almost all commentators put much of the funding success down to Everitt's undeniable political skill. Typical of this, writing in the Stanford Today (March/April 1997 edition) Robert Lee Hotz says "The project has been threatened with cancellation seven times - more times than almost any other project in the agency's history, NASA officials acknowledge. Yet it has survived where many other more widely supported projects have been scratched, due in no small measure to Everitt's political acumen." http://www.stanford.edu/dept/news/stanfordtoday/ed/9703/9703fea1.html accessed on 25/08/2016.

[17] The website claims that fully nine new technologies were developed during the lifetime of GP-B.

[18] https://www.sciencenews.org/article/gravity-probe-b-finally-pays accessed on 25/08/2016.

[19] Although note the reference I make, in Section 3.2 above, to the idea pondered by Blackett as early as the 1930s.

be seen as passing through 4 or 5 distinct stages; although it was not necessary to complete one before commencing the next. I would say that the first stage, before moving fully into the design and production stages—followed by the actual mission stage and then the post-mission data analysis (and publication) stage—included the initial conception and idea for the project. Together with initial research and development, this ideation stage also involved communicating the ideas as broadly as possible and drumming up support for the project. This then led into the stage of instrument design, starting in 1964 when NASA funding first arrived; indeed Will, 2015, p. 1, tells us that the GP-B project did not start until NASA funding for it was agreed in 1963. Thus, once the initial idea had won support and the project appeared feasible, the science instruments that were to be used in the experiment had to be designed. It was at this stage that the scientists involved in the project actually had to decide what they were going to measure and exactly how they proposed to do that. Although the two effects they hoped to measure had been theoretically established long before (in 1916 and 1918, as I have said), this project represented the first time that anyone had set out to measure either of them.

I will now try to explain each of the effects in turn, and the combination of the two that GP-B aimed to measure. Although as I have already mentioned and as we will see in detail in Chapters 5 and 6, the post-flight concerns and the emphasis of the data analysis shifted, it is vital to understand clearly what the objective of GP-B always was. I will start by explaining the larger of the two effects and then move on to the combination of this de Sitter effect with the Lense-Thirring effect of frame dragging. The geodetic effect is much the larger of the two effects GP-B set out to quantify and therefore easier to detect and measure.[20] Figure 3.1 is a schematic representation of how the two effects were expected to alter the direction of the spin axis of the GP-B gyroscopes.

[20]For GP-B 6,606 compared to 39 mas.

3.3.1 The geodetic effect

Einstein's GTR describes gravity as a correlation[21] between the presence of energy–momentum and a warping of the 4-dimensional spacetime away from a flat (Euclidean) continuum to a variably curved spacetime. The geodetic effect was very quickly identified by de Sitter as a consequence of this warping of spacetime around all massive bodies. To gain an idea of how the effect is produced, it is best to consider the operation known as parallel transport and how this can be used to examine and reveal curvature, which I now explain.

The idea of parallel transport, which is very closely related to the geodetic effect, provides a way of probing the curvature of spacetime. To see how parallel transport works we need to consider a vector defined at some point on a curve. Parallel transport consists of moving the vector along the curve in infinitesimal steps, maintaining—as the name indicates—each new vector parallel to the previous one. Clearly, we can do this around a closed path consisting of any arrangement of curves. In a flat (hyper)space, when we return to the starting point, the vector which has been parallel transported around the closed path will always be pointing in the same direction as the initial vector. However, this is not necessarily the case for a curved (hyper)space and thus it does not necessarily hold for curved 4-dimensional spacetime, depending on the path chosen.

To see this, we can consider a closed path on the surface of a sphere.[22] The sphere is a useful system for making comparisons with curved spacetime as

[21] It is very tempting (and quite common) to say that the warping is caused by the presence of a massive object, but it is important to avoid such a formulation as it is quite clear that GTR makes no such claim about any causal relationship and that we would be equally justified in claiming that the warping of spacetime causes the characteristic properties of mass. The only statement we can correctly make is that there is a definite correlation between the two.

[22] I explain this here in the case of a 2-dimensional surface embedded in 3-dimensional space, as it is the easiest to visualise. A sphere is a 2-dimensional surface with constant positive intrinsic curvature, but the result of parallel transport that I illustrate can be generalised to more dimensions.

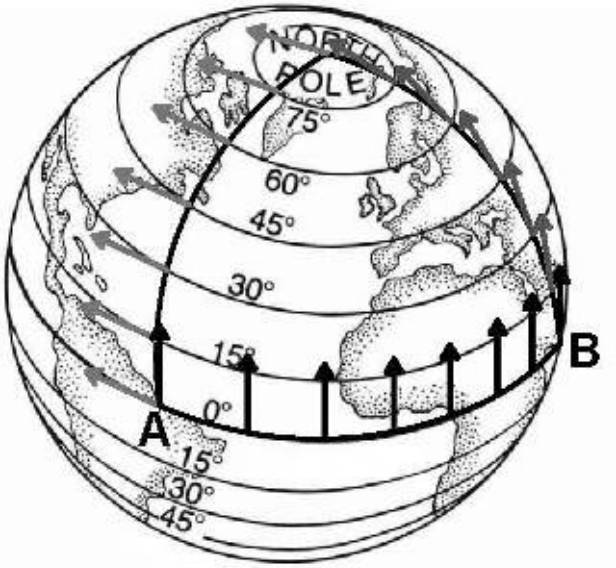

FIGURE 3.2: Parallel transport around a closed path on the Earth of a vector initially pointing towards the North Pole (NP). First it moves from A to B (black arrows); then (mid grey arrows) to NP; and finally (light grey) it returns to A. After this closed loop, the orientation of the vector has shifted by $90\,^{\circ}$.

at every point on a (large enough) sphere, such as the Earth,[23] the surface appears locally flat. It is only by minute local experimentation, travelling over large areas of the surface and using reference points that are not on the surface itself or removing oneself from the surface of the sphere that the overall curvature becomes apparent. Somewhat counter-intuitively, parallel transport may result in a change in the direction of the original vector when it returns to its starting place; but the change depends on the specific path followed.

For example, we could follow a closed path formed of segments of three circumferences of a sphere, as if they were great circles around the surface of the Earth, as illustrated in Figure 3.2.[24]

[23] As we know, the Earth is not a prefect sphere and to get a better idea here we also need to ignore the local orography, so maybe it is best to consider the surface of a (completely calm) ocean for the purposes of the example.

[24] My example follows that in Schutz, 1985, pp. 164-5.

The first segment runs from a point on the equator, A, which in the example illustrated in Figure 3.2 is just off the coast of Brazil, near the mouth of the Amazon, due east of where the equator intersects the east coast of Africa: point B. We can consider a straight vector at A, tangent to the surface of the Earth, initially pointing due north: towards the North Pole (NP). We then parallel transport this vector along the curve followed by the equator from A to B. That is, at each successive infinitesimal step along the path a similar vector is drawn parallel to the previous one, thus treating the globe as locally flat. A few such intermediate steps are illustrated in the figure to show how the vector would move along this path. At B, the vector is still pointing due north to NP: with each infinitesimal step, the same direction is maintained. The second segment now runs from point B, due north to NP. As this path follows the direction in which the vector is now pointing, by maintaining the same direction with each successive infinitesimal step (parallel transporting the vector as we go, so it is always pointing tangentially along the Earth's surface) the vector hugs, so to speak, the curve of the Earth and arrives at the NP pointing (locally) back down the other side of the globe towards the equator, at a point in the middle of the Pacific Ocean. From here, NP, mirroring the movement along the first segment, we can now close the path by parallel transporting the vector due south back to A. Once again, Figure 3.2 shows a few of the steps along this segment of the closed path. With each infinitesimal step around the globe, the vector continues to point towards the same location in the middle of the Pacific Ocean, just as it pointed to NP during the whole of its parallel transport along the first segment. It should now be clear that when the vector arrives back at its starting point, A, it is pointing in a different direction: in the case illustrated, at a $90\,^\circ$ angle.

This demonstrates the curvature of the surface on which we are working, since on a flat surface, the vector always points in the same direction, by the definition of parallel transport, so when it returns to its starting position it has to be pointing in that same direction. This effect was exploited by Riemann and led him to define what we now know as the Riemann tensor, which defines the intrinsic curvature of an n-dimension space. As a property of 4-dimensional spacetime, the Riemann tensor is defined in terms of the metric $g_{\mu\nu}$ (and its second derivatives). The definition of a flat spacetime is $R^{\alpha}_{\beta\mu\nu} = 0$.

That is , the Riemann tensor vanishes everywhere on a flat manifold. In his comprehensive introduction to GTR which has now become a classic text, Schutz tells us that:

> $R^{\alpha}_{\beta\mu\nu}$ must be the components of the $\left(\frac{1}{3}\right)$ tensor which gives δV^{α}, the component of the change in $\vec{V}$ on parallel transport around a loop.
>
> (Schutz, 1985, p. 169)

It was by contracting the Riemann tensor and combining the result with the metric, $g_{\mu\nu}$, that Einstein defined the symmetric tensor $G^{\alpha\beta} \equiv R^{\alpha\beta} - \frac{1}{2}g_{\mu\nu}R$, which is instrumental in describing gravity and forms the basis of his field equations (Equation 2.1).

From the example of parallel transport on a sphere, it should be clear that the change in the direction of the vector depends on the path; had we continued all the way around the globe following the equator, the vector would have returned to A pointing in the original direction: towards NP. It should also be clear that the effect is cumulative and that if we trace out the same path many times we will accumulate a shift in the direction of the vector.

The geodetic effect results from spacetime curvature; for a spinning body in a 4-dimensional spacetime, the effect is similar to the parallel transport of a tangent vector on a curved surface. Of course, that we are aware of,[25] no object can actually follow a closed spacetime path: to do so it would either have to travel "instantly through space", return to its original position at the same time it left; or it would have to travel "backwards through time", exiting what we think of as its forward light cone and "doubling back" to its starting event. In the general case, the intrinsic spin vector of the body is not parallel transported around the path, it is in fact Fermi-Walker transported. However, if the acceleration of the body is zero, that is to say, if it is in free fall and follows a geodesic of the local spacetime, then we recover parallel transport. In this way it is possible to demonstrate the curvature of local spacetime, or the warping of spacetime away from the flat, by transporting a spinning body along a geodesic of the curved spacetime. Since in the vicinity of the Earth

[25] Although, certain quantum effects do seem to suggest that information at least can move instantaneously; on some readings, anyway.

the curvature of spacetime is extremely small, the change in the direction of the spin vector is minuscule. To make this change appreciable, GP-B exploits the cumulative nature of the effect. De Sitter calculated the effect for the Earth–Moon system (effectively a spinning body) orbiting the Sun; it was empirically confirmed in 1987[26].

3.3.2 Frame dragging and combining the 2 effects

Measuring the frame-dragging effect is considered by many to have been the main objective of GP-B, since it had never been directly measured before the experiment.[27] As I indicate at the start of this section, when GP-B was first conceived, neither had the geodetic effect been measured; but as I have just stated, it was successfully measured in 1987. This was due to the situation having changed after Apollo 11 landed on the moon and the crew placed a large reflector on the Moon's surface, later to be joined by others. Those reflectors have since been used in what are known as lunar ranging experiments. By bouncing laser beams off the reflectors it has been possible to acquire extremely accurate information about the position and orbit of the Moon and this yielded the first empirical confirmation of the de Sitter effect, many years before GP-B was launched. This has led to the extra importance given by many to the frame-dragging effect after 1987.

[26] See my previous footnote in Chapter 2 concerning this: footnote 29

[27] Throughout almost the entire history of the project the effect had not been detected at all. However, Ciufolini et al., 1998 announced they had succeeded in confirming the effect predicted by GTR to within 20% accuracy using data from the two LAGEOS satellites. The GP-B line in response became that the effect had still not been directly measured, although it may have been detected. Returning to the LAGEOS data and combining it with new satellite data concerning the Earth's multipole moments, Ciufolini and Pavlis, 2004 announced that they had succeeded in detecting the effect with an accuracy of approximately 10%. As this was the year when GP-B was finally launched, the announcement had no effect on the project at all, although it clearly affects the status of one of its fundamental goals. However, the additional results from Ciufolini et al. may actually provide valuable support for GP-B if they are in agreement. Both the 1998 and the 2004 results agreed with the GTR prediction of frame dragging.

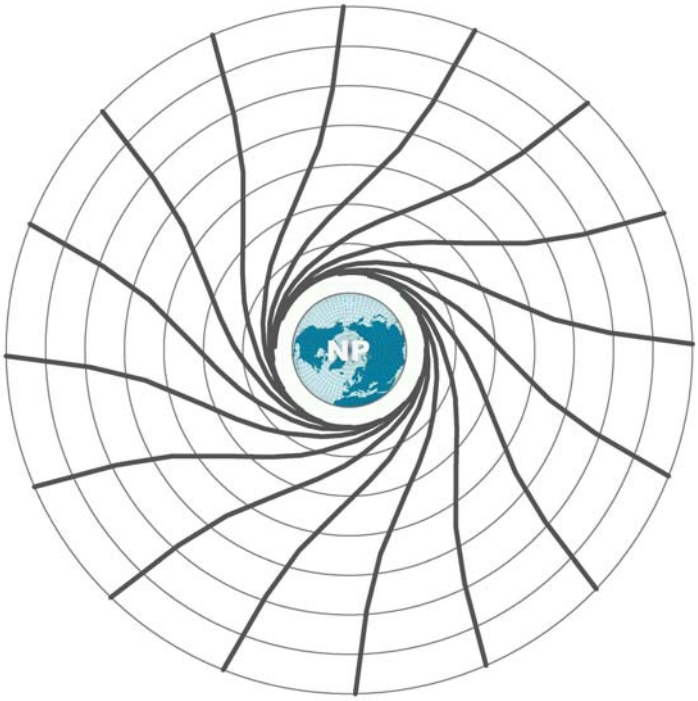

FIGURE 3.3: Gravitational attraction is dragged around and twisted as the central gravitating body (the Earth) spins.

Even working only to the accuracy of a post-Newtonian approximation, any GTR metric with a spinning mass at its centre gives rise, in addition to the geodetic effect, to a rotational frame-dragging effect, which in contrast to the former is not present around a non-spinning mass. Figure 3.3 is a representation of the way in which we can think of spacetime being dragged around by a central spinning massive body: the Earth in this case. This effect of spin had been of interest to Ernst Mach, acknowledged by Einstein as one of the people who most influenced his early work.[28] Mach had considered the effects of spin in his attempts to fathom the relation between the distant "fixed" stars and local inertia, using a spinning outer shell to represent the (relative situation of the rest of the) universe around a central observer. Einstein had taken

[28]For example, in Autobiographical Notes of the celebrated volume edited by Paul Schilpp, Einstein tells us that Mach's History of Mechanics "exercised a profound influence on me in this regard when I was a student"; that regard being precisely Mach's rejection of "dogmatic faith" in classical mechanics as the "firm and final foundation of all physics". (Schilpp and Einstein, 1998[1949], p. 21)

this up within his attempts to divest spacetime of any absolute component and had worked on the corresponding metric field of a rotating Minkowski spacetime (Hoefer, 1994). Indeed, it seems that precisely the failure of such a metric to give a solution to the equations in his draft 1913 "*Entwurf*" theory was one of the crucial steps leading to the final GTR in 1915. However, it seems that while Einstein had worked on different calculations of possible frame dragging, both within and outside a rotating shell, it was only in 1918 that Lense and Thirring published the first full account of the phenomenon outside a spinning body.[29]. Hence it is credited to those authors and can be considered in some sense as an extension of Einstein's GTR, though certainly not a consequence Einstein was unaware of.

The model usually used in GTR for a spinning central body is the Kerr metric, but other well-known solutions such as the Neugebauer-Meinel metric for a self-gravitating rotating disk or the Kasner metric, together with all rotating source solutions, also produce frame dragging.[30] The fact that the effect appears in different solutions of EFE makes it extremely unlikely that it is simply a consequence of the individual models; rather it must be considered a consequence of Einstein's theory itself (not of a particular solution of his field equations). Although the frame-dragging effect was not explicitly stated by Einstein, it is certainly well within the realm of his theory, as a description of gravitational effects; and it is closely related to the importance Einstein gave to the "Machian" idea that it is only energy–momentum that determines the spacetime metric and thereby the motion of freely moving bodies.[31]

[29] In this respect, when recounting "Einstein's Quest for General Relativity" in the Cambridge Companion to Einstein, Michel Janssen and Lehner, 2014, p. 198; italics in the original, tell us of the letters between Einstein and Thirring in 1917 and go on to conclude:

> Following up on his study of the effect of a rotating hollow shell on the metric field *inside* of it, Thirring studied the effect of a rotating solid sphere on the metric *outside* of it [Lense and Thirring 1918]. Einstein [1913c, 1261] had also pioneered calculations of this effect, now known as "frame dragging".

[30] This result is stated by Collas and Klein, 2004, p. 1197.

[31] For example, when discussing Einstein's motivation for introducing a cosmological constant into his field equations in 1917, (Torretti, 2000, p. 177) tells us that

Of course, prior to empirical confirmation, there existed the possibility that what Lense and Thirring had published and interpreted as a physical effect was in fact superfluous structure within the mathematical expression of GTR, with no observational consequences. It could conceivably have been the case, for example, that the calculations that Lense and Thirring published mistakenly contained a mere coordination effect that actually had no observable consequence.[32] While there was no empirical confirmation of the effect, frame dragging was merely a theoretical conjecture open to being demonstrated or refuted (though as I say above, not in a way that Einstein could foresee, due to the tiny magnitude of the effect within the Solar System).

The first calculation of this effect published by Lense and Thirring was performed in the context of a linearised analysis where the metric is expressed using a scalar potential and a 3-vector potential.[33] In the linearised approach they adopted, the magnitude of the rotational effect can be seen as being a combination of two separate contributions: one from the warping of space and the other from the dilation of time. The overall effect was interpreted as a twisting of the spacetime fabric in the vicinity of a spinning massive body; and this is precisely what GP-B aimed to measure, to an accuracy of 1%.

Einstein was unsatisfied with the separate condition of flatness at infinity of the Schwarzschild solution and that Einstein thought it was:

> in crass violation of a principle that Einstein attributed to Mach and regarded as one of the groundstones of general relativity. Einstein Einstein (1918, p. 241) stated this "Machian Principle" (Mach'sches Prinzip) as follows: "The [metric] field is *exhaustively* determined by the masses of bodies."

[32] In a similar way, we can see, for example, that EFE allow for a solution consisting of an empty universe except for some curvature associated with the metric, or gravitation of spacetime itself. But identification of such a possible solution in no way leads to any additional observational consequences: it is simply a mathematical option within the structure of the theory. It can, nonetheless, be used to argue that we should adopt a substantivalist notion of spacetime. It can be argued that if there could be curvature in the absence of energy-momentum, then spacetime itself must be responsible and therefore it should be considered as being a possible recipient of properties.

[33] This is reported in Thorne, 1988, p. 577. See also Misner, Thorne, and Wheeler, 1973, Chapter 18, for an explanation of linearised GTR.

Seeing as the 2 effects that GP-B set out to measure result from different properties of the Earth (the central gravitating mass itself, in the case of the geodetic effect; and the spin of that central mass, in the space of frame dragging) it was quite straightforward—and was contained in the original idea—to produce a set-up in which the two effects would cause perturbations in the intrinsic spin of a test gyro that were at right angles to each other. This was achieved simply by using a gyro in a polar orbit: orthogonal to the intrinsic west–east spin of the Earth. The problem which was cited as the most intractable, was how to measure the orientation of the spin axis of a gyro in free fall without interfering with its motion. It was the London moment (LM) that offered a solution to this problem and that was ultimately what GP-B actually measured; which is what I describe in detail in the next section.

3.4 Instrumentation: What to Measure and How

Once enough people at Stanford had been convinced that the scientific project was viable, experimental design was started and it was necessary to develop and test both a spacecraft that would be the satellite placed into orbit around the Earth in which the entire experiment was to be performed, and the science instrument assembly (SIA). The latter would contain the key elements of the experiment, and needed to be developed together with the on-board systems that would provide the necessary support for the scientific experiment. As I indicate in the introduction to this chapter in Section 3.1 above, I will not go into the details of these engineering and technical aspects which are beyond the scope of the work I present in this thesis; but in Chapter 5 and Chapter 6, I do consider some of the claims made concerning the performance of the science instruments.

As I have mentioned, the fundamental ideas on which the project was based did not change after they were originally brought together in the early 1960s. They consisted of using a superconducting sphere as the gyroscope at the centre of the experiment. In his extremely brief original 1960 publication, Schiff simply says: "A secular precession of 6×10^{-9} radian per day would be very difficult, but perhaps not impossible, to observe ....[with] the possibility of

using for this purpose a gyroscope that consists of a superconducting sphere" (Schiff, 1960). However, the technical aspects of how to meet the challenges and resolve the problems that would arise throughout the history of the experiment were anything but clear at that initial point. Moreover, exactly how to measure such an "angular rate ... to an accuracy better than 0.5 marcsec/year" was something that the original proposals did not contemplate and the technology to do so just was not available at the time. Schiff simply muses of the possibilities if such a system "could be made to operate exceedingly well" (Schiff, 1960, p. 216).

Although I describe the basic idea of the experiment as straightforward, of course that certainly does not mean that it can in any way be seen as simple. It is hard for us, heading rapidly towards the third decade of the 21st century, to imagine the scale of the challenges that the team were faced with both initially in the ideation stage of the experiment, and through much of the actual experimental design. Let me just recap the main technical issues that required solutions that went way beyond the fact that we could successfully place an artificial satellite in orbit and receive signals from it here on Earth (which was why the project seemed even remotely possible at the time; but it was also just about as far as the relevant technology went when Schiff first published his idea!).

- It was necessary to make a near-perfect gyro (improving performance by a factor of 10^5 compared to the best electrically suspended gyroscope on Earth[34]) with a coating that would be a superconductor at the experimental temperature, and of course to maintain that temperature within the satellite.
- The gyro had to be spun up, aligned with an appropriate guide star and then released into a purely gravitational near-Earth orbit in which it would maintain its alignment to such a high degree of accuracy that the predicted relativity drift, or any significant variation from it, could be detected.

[34]In the paper of the 2015 Classical and Quantum Gravity Focus Issue on GP-B that specifically deals with the gyroscopes, we are told: "The measurement required a factor of 10^5 improvement in the state-of-the-art of conventional gyro drift rate error."Buchman et al., 2015, p. 1

- The satellite and all the experimental instrumentation it contained had to follow the path of the freely falling gyro perfectly, so that the on-board equipment could monitor the experiment, virtually without interfering with the gyro, and send the read-out to Earth.
- It was necessary for the satellite to contain a telescope that could determine the direction to the specified guide star with unprecedented accuracy and somehow provide that direction as the baseline, so to speak, against which measurement would be made.
- The proper motion of the guide star had to monitored, so that the actual direction in which the telescope indicated the guide star could be compared to inertial space, via the "fixed" background of distant stars (assuming that these would all be too faint to serve as an appropriate guide star).
- Finally, and maybe the greatest conceptual challenge that the idea of the test presented, the orientation of the gyro had to be determined without affecting its functioning; this then had to be compared infinitesimally with the direction to the guide star determined by the telescope.

These were the basic requirements in order to be able to perform the experiment that Schiff had originally envisaged that would allow the two effects of gravitation to be detected and quantified over the roughly one year duration that the experiment was planned to have. From calculations of the precession that GTR predicted the gyro would experience while orbiting the Earth, these general requirements were refined to give the following 4 technical goals:

- a near-perfect torque-free gyro with a drift rate $< 10^{-11}$ degree/hour
- a satellite-mounted telescope capable of tracking a star to an accuracy of better than 0.5 mas
- knowledge of the guide star's "real" position during the entire experiment accurate to at least 0.5 mas/year
- spacecraft orbital position information sufficiently accurate to calibrate the gyro read-out with an accuracy that would leave the relativity signals resolvable

The first of these was dictated by the size of the predicted GTR effects and required that the gyro be almost perfectly spherical, have an almost precisely homogeneous mass distribution and that any charge on the surface be evenly spread across it. The other three were necessary if the "baseline" of the experiment (the original direction in which the gyro spin axis was pointing in inertial space) was to be known and the change with respect to it monitored sufficiently accurately over the year-long experiment for the drift due to relativity to be observed. These 4 basic goals led to what are sometimes described as the 7 "near zeroes" or fundamental design requirements for GP-B to succeed. The seven are:

- Electric dipole moment of the rotor < 0.1 V·m which is the same as a dipole equivalent field of trapped flux in rotor < 9 μG
- Rotor asphericity < 10 nm
- Centre of mass of the rotors < 50 nm from geometric centre
- Rotor electric charge < 108 electrons (or equivalently < 15 pC)
- Magnetic field strength at rotors $< 10^{-6}$ g (attenuation of ambient fields $< 10^{-12}$)
- Gas pressure $< 10^{-12}$ torr
- Cross track acceleration (transverse to roll axis) $< 10^{-11}$ g

These do not include the accuracy and knowledge of the spacecraft and telescope pointing, which are sometimes added to the list: spacecraft pointing error < 20 mas; telescope pointing knowledge during the guide star valid (GSV) period[35] < 0.1 mas. Furthermore, the temperature of ~ 2.7 K, which among other things contributes to maintaining the noise in the gyro and telescope read-out signals near zero by reducing thermal white noise, can

[35]With the telescope and the entire SIA aligned with the guide star, which to give rise to the two relativity effects in orthogonal planes (roughly north-south and east-wet) needed to be located in the plane of the polar orbit of the satellite, the SIA would be occluded behind the Earth for a substantial part of each orbit. Primary data collection was thus to take place only during that section of each orbit during which the guide star was visible from the satellite and the on-board telescope had successfully acquired it. This was what was referred to as the "guide star valid" (GSV) period of each orbit.

Rotor Properties	
Density homogeneity	$< 6 \cdot 10^7$
Sphericity	< 10 nm
Electric dipole moment	< 0.1 V·m
Environment	
Cross track acceleration	$< 10^{-11}$ g
Gas pressure	$< 10^{-12}$ torr
Magnetic field	$< 10^{-6}$ gauss
Mixed	
Rotor electric charge	$< 10^8$ electrons

TABLE 3.1:
Seven "near-zero" GP-B parameters required for a successful mission.
[SOURCE: Based on Table 1 in Kahn, 2008, p. 12]

be included as another "near zero," leading to ten—or more. In fact, when reviewing these requirements after the mission and the results had been published, Everitt et al. tell us that:

> GP-B hinged on nine essentials . . . : three cryogenic, three met by the low-g of space, and three by spacecraft roll. These led to 12 fundamental requirements defining management and instrument layout.
>
> (Everitt et al., 2015, p. 4)

Establishing experimental requirements is one thing, but it is a very different matter to actually design an experiment to meet them and then implement that design. It took the Stanford team and their partners over 40 years to arrive at a position where they could claim to have met them all. Some requirements can be seen as more closely related to the design and set-up of the experimental hardware; while others are more closely linked to the data processing and analysis side of the experiment. However, hardware on the one hand, and data processing and analysis on the other are certainly not two independent parts of an experiment: they are usually closely linked branches of a whole, as was certainly the case with GP-B.

In any experimentation in physics, depending on the options available and the convenience of tackling obstacles in one fashion or another, effects can be physically shielded against, sensitivity can be increased, or data handling can be refined to remove predicted signals and effects that are not the target of the experiment. Back in 1989, in her book which she introduces by affirming: "Science is measurement; capacities can be measured; and science cannot be understood without them" Nancy Cartwright, 1989, p. 66, cites GP-B as an exceptional experiment which aimed "not to calculate disturbing effects but to eliminate them". As we will see in Chapter 5 and Chapter 6, in the end both methods were needed to arrive at final results; but as Cartwright explains, that was not the initial intention.[36] There is a further possibility that became a part of the ongoing and long-lived GP-B project, which is to make use of (or require there to be!) improvements in other areas of research, and incorporate the results of such advances into the project. All these aspects are common to many modern physics experiments and all were employed or adopted throughout the lifetime of GP-B.

On different occasions, in an example of the methods of experimental science at their best, the GP-B team were able to turn adversity to their advantage. They managed to overcome what initially appeared to be insurmountable technical problems resulting from the theory, and convert them into useful—maybe even indispensable—parts of the experiment. This is the case with the London moment (LM) produced by a rotating superconductor, such as the gyro proposed by Schiff. The effect had been predicted in 1948[37] but was not detected until 1963—by the GP-B team, among others. An LM is a magnetic moment that accompanies any spinning superconductor and is directed precisely along the spin axis of the superconductor. The effect meant that the superconducting surface of the gyroscope would cause a large perturbation in electromagnetic measurements. This appeared to be a major

[36]The front cover of the first edition of Cartwright's book was a diagram of the GP-B dewar containing the experimental apparatus; an indication of the milestone the experiment was seen as in the history of scientific experimentation and the measuring of "nature's capacities".

[37]For details of the way in which Fritz and Heinz London used a totally innovative idea to develop a new model within electrodynamics which allowed them to account for the Meissner effect, see: Suárez, 1999.

problem until it was realised that the effect could actually be used as the principal marker of gyroscope spin axis orientation. Indeed, this is what the experimental set-up could best measure to determine the original spin axis orientation of the gyro and the change in that orientation over the duration of the experiment; always while causing an absolute minimum of interference to the gyro itself.

So the experiment was designed to measure the LM, which coincided with the gyro spin axis orientation. This was to be achieved by placing an extremely sensitive magnetometer in the housing around the gyro. This would then produce the principal experimental read-out. But for a magnetic moment to produce a reading, the pick-up loop of the magnetometer has to move through or across the resultant field. With the gyro orientation set not to move (by more than the order of the predicted relativity effects over the year-long experiment) the solution was to move the pick-up loop around the gyro. Thus a fixed pick-up loop was built into the gyro housing and the entire housing and satellite would roll around the gyro at its centre. This rolling of satellites is an extremely common and useful technique for all sorts of technical reasons that can be summarised as the effect of evening out perturbing effects over each complete roll. Of course, for this to work and the on-board telescope to track the guide star accurately, the roll axis had to coincide almost exactly with the on-board telescope line of sight.

In this way the basis of the design of the science instrument assembly was settled. It can be seen in Figure 3.4. The entire SIA, mounted on the body of the satellite, would roll around its longitudinal axis, coinciding with the telescope line of sight and thus the direction to the guide star. The magnetometer pick-up loop fixed in the gyro housing that constituted the body of the SIA would then roll around the gyro and its LM, thereby producing the science signal. That signal would thus vary in magnitude at the roll frequency of the spacecraft as the pick-up loop orientation moved around the fixed gyro spin axis orientation that coincided with the LM. In this way the direction to the guide star would be given by the orientation of the SIA itself; while the relative orientation of the gyro, compared to the SIA housing, was given by measuring the LM produced as the housing rolled around the gyro.

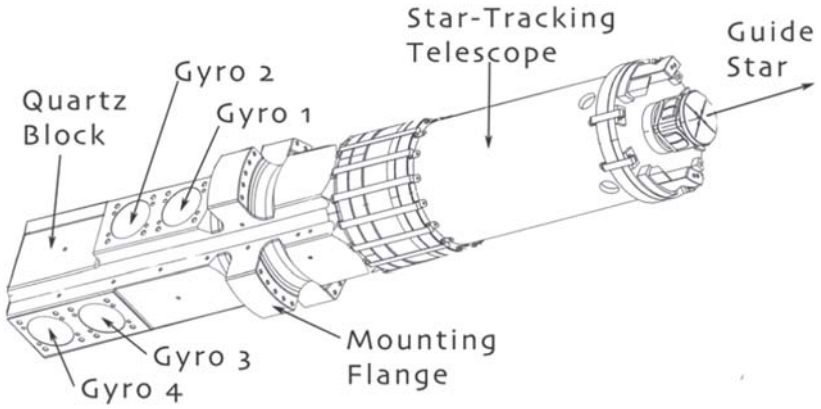

FIGURE 3.4: Representation of the SIA showing the location of the 4 gyros within it and the relative position of the telescope. [SOURCE: Everitt et al., 2015]

The magnetometer to be used in GP-B to compare the orientations of the two basic elements—gyroscope rotor and telescope—in the way described was a superconducting quantum interference device (SQUID). The SQUID was invented in 1964, just as serious design work on GP-B was getting underway and was incorporated into the design.[38] This was the basic idea behind the experimental design. It is illustrated in Figure 3.5. As the angle, represented by θ in the figure, between the orientation of the gyro spin axis (fixed, except for the relativity effects[39]) and the satellite roll axis (which coincided with the

[38]The working principle behind a SQUID is that a (tiny) current flowing around a superconducting loop will be affected by an external magnetic field; as would any electrical current. In the case of a SQUID, however, the effect can be detected for the tiniest of external magnetic fields, making it the most sensitive type of magnetometer we know. The set-up of a SQUID is such that the current induced in the superconducting loop by the external magnetic field represents an addition to the current in one branch of the loop and a subtraction from the current in the other. The branches contain specific (Josephson) junctions which produce a voltage when a certain critical value of the induced current is reached. It thus converts the external magnetic field into a voltage reading.

[39]In fact, there were also tiny additional effects predicted to be present in the read-out signal which could be calculated prior to the experiment and subtracted from the science signal. These were a correction due to the actual oblateness of the geoid, compared to the idealised sphere used in the calculations of the geodetic effect; the tiny degree of eccentricity on the GP-B satellite orbit (Silbergleit et al., 2015a, p. 6); the solar geodetic effect that results from the Earth's orbit around the Sun (Conklin

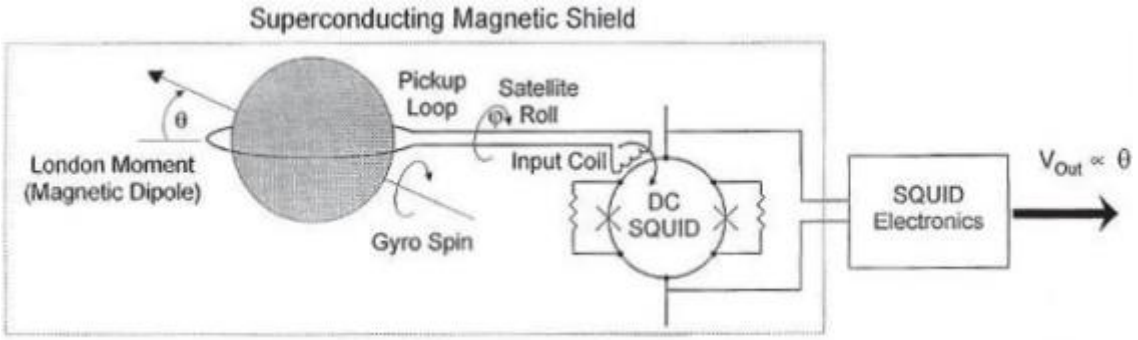

FIGURE 3.5: Representation of the London moment that is generated by a spinning superconductor and aligned precisely with its spin axis. This is the effect that was used to determine the orientation of the gyro spin axis in GP-B.
[SOURCE: Muhlfelder et al., 2015]

direction to the guide star through also being the telescope pointing direction) varied, so the voltage produced by the SQUID as the science output signal varied.

In the same way as the LM was initially seen as a drawback but the team turned it to their advantage, the aberration of starlight caused by the motion of the telescope (both in orbit around the Earth and in orbit with the Earth around the Sun) was likewise thought to be an awkward side effect that would have to be compensated for. There is no way of avoiding this effect, so it always has to be taken into account, but—once the exact value of the effect had been calculated—it too proved to have a useful application. It was used as a means of calibrating[40] the read-out from the most important on-board systems with every orbit of the satellite around the earth (also over the entire year of the Science Phase, using the annual aberration). The expected read-out data for a typical GSV period is shown in Figure 3.6. The effects of orbital aberration can clearly be seen as amplitude of the signal grows from its initial value, when the satellite is heading almost directly towards the guide star, to

et al., 2015, p. 46); and also, gravitational bending of the light reaching the satellite as it passed the Sun (Everitt et al., 2015, p. 4)

[40]Or "recalibrating" as initial calibration was performed before the Science Phase of the experiment started, during an initial calibration phase, and so during the Science Phase this was an additional step to check and fine-tune that initial calibration.

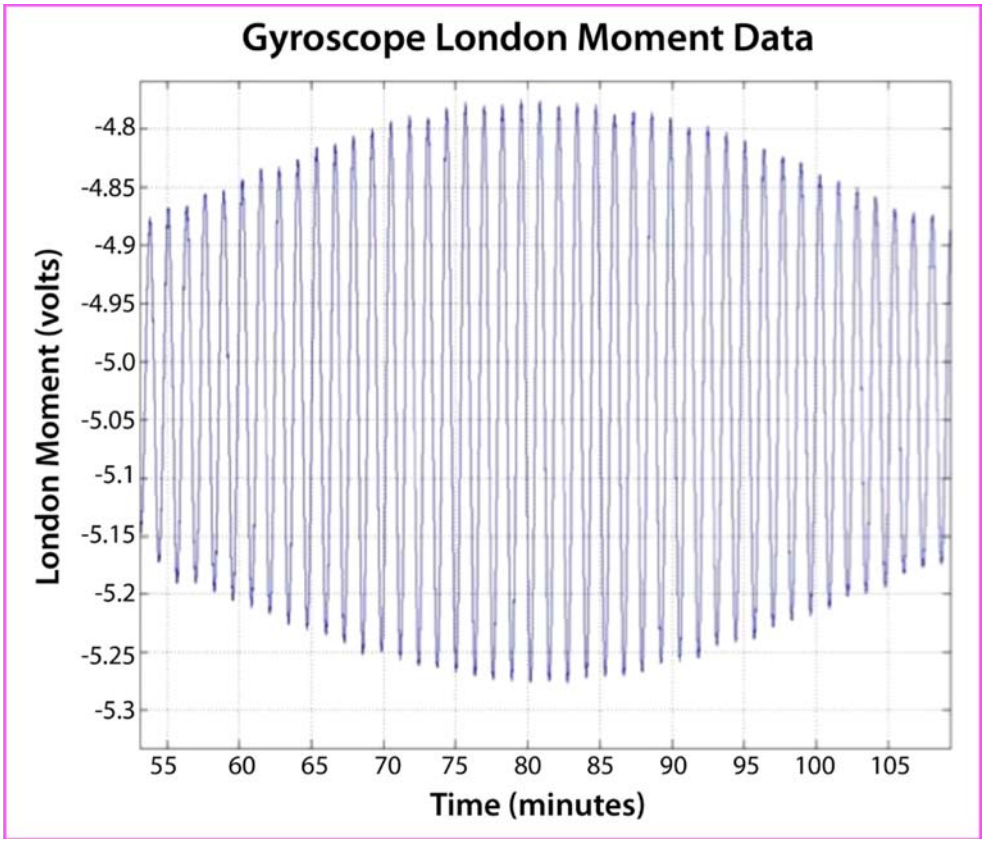

FIGURE 3.6: The figure shows the GP-B readout science data as it was expected. As the SQUID pick-up loop rolls around the gyro, the gyro LM produces a voltage signal in the SQUID circuitry. This signal varies at the roll rate of the space craft (completing one full sine wave every 77.5 seconds, as can be seen). Over the duration of the GSV period, lasting just under 60 minutes of every 98-minute orbit, the signal is modulated by orbital aberration, which is responsible for the outer envelope around the continuous sine waves. To calculate the drift over the duration of the experiment, the data from each individual GSV period—as shown here—was to be combined into one data point. When combined with the detailed information of the telescope pointing direction and the exact position of the guide star at that time, it would represent the gyro spin axis orientation for that orbit and could be compared with the results of every other GSV period to give the drift, both east-west and north-south separately.
[SOURCE: Everitt et al., 2015]

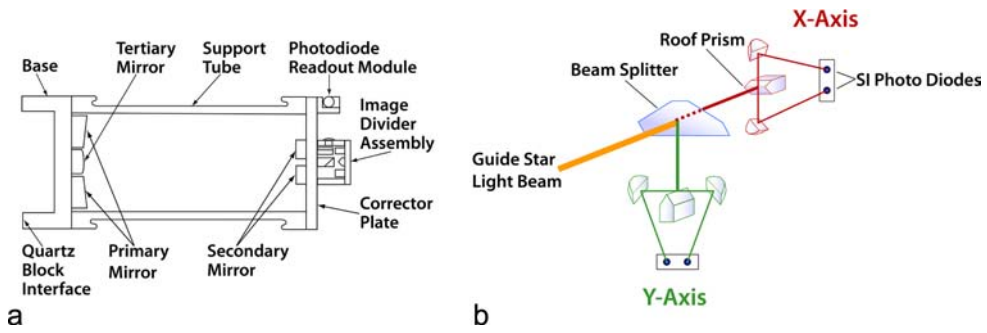

FIGURE 3.7: In **a**, the arrangement of the different components of the GP-B on-board Cassegrain-Schmidt-type optical reflector telescope are shown. Then **b** represents the way in which the incoming light beam from the guide star was split into two perpendicular planes and then the image was centred in each plane individually by ensuring that the intensity of the incoming light was the same in each half of the image.
[SOURCE: Everitt et al., 2015]

its maximum value half way through the GSV period of the orbit (when the satellite is travelling perpendicular to the line of sight to the guide star) and returns to its initial value again at the end of the GSV period, just before it is occulted behind the Earth (again, travelling directly away from the guide star). The data from each individual GSV period can thus be averaged to provide the gyro axis orientation for each orbit. It is necessary to calculate the scale factor to convert the voltage output signal into an angle: the output is proportional to the angle between the telescope line of sight and the gyro spin axis orientation: θ in Figure 3.5. This can be calculated from the orbit aberration value, as well as from the data gathered during the initial orbit checkout (IOC) phase.

The components of the Cassegrain-Schmidt-type optical reflector telescope that formed part of the SIA were made of fused quartz (see) to proportion them with the rigidity and strength required for the precision demanded of it. The telescope was required to track the centre of the guide star to within 0.1 mas (compared to the predicted cumulative frame-dragging effect of 39 mas over a year) (Everitt et al., 2015). That was a degree of accuracy which had never before been reached by a telescope. It was achieved through a

system of splitting the incoming light beam from the guide star into two halves and using sensors to measure separately the intensity of the light signal from each of the two respective halves of the star. When the intensity was precisely the same from the two halves of the image (in each of 2 perpendicular planes) it was assumed[41] that the telescope was precisely centred on the star. A schematic representation of the telescope used in GP-B is shown in Figure 3.7 which shows the overall arrangement of the assembly and how the incoming light beam from the guide star was split and the image centred in two perpendicular planes.

The guide star used in the experiment was IM Pegasi (HR 8703) whose magnitude, right ascension and proximity to a quasar representing the "distant fixed stars" together with the fact that it is a radio source, all made it suitable. The on-board telescope locked onto the guide star whose position relative to a nearby fixed, background quasar was then monitored using Earth-based very-long-baseline interferometry (VLBI) radio astronomy techniques. Since the telescope and the SIA were both made of fused quartz, they were bonded together to form one complete whole, with the only (absolutely minimal) distortions of shape and size of the equipment (which would have led to inaccuracies in the measurements) being those of the whole crystalline quartz structure itself.[42] Furthermore, to meet the requirements given in Table 3.1, the experiment was performed in an extremely high-level vacuum to avoid particles colliding with (and thereby interfering with the motion of) the gyroscope rotor, or any of the other equipment; the SIA was maintained close to liquid helium temperatures (2.7 K); it was shielded from electromagnetic radiation that could affect the gyroscope; and the experiment was performed in an environment with an extremely low magnetic flux. Once both the scientific and the space-technology aspects of the experiment had been developed, the

[41] There are, however, possible sources of error introduced into this method from any object, such as interstellar dust, that momentarily (or systematically) obscures one side of the star, or from a lack of symmetry in the luminosity of the star itself due to a tendency to emit solar flares from one or other pole, for example. Despite my worries concerning such possible sources of error, they do not appear to be mentioned in the literature concerning GP-B at all; so I must assume that they were taken into account during the selection of the actual guide star used in the experiment.

[42] This is one of the many industrial spin-off success stories cited on the GP-B web site that I mention in an earlier footnote in this chapter (footnote 17).

spacecraft had to be launched, and the data collected and analysed. Clearly it is a question of interpretation exactly what the importance of GP-B was expected to be. Heifetz et al., in a style seemingly befitting their role as data analysts, provide a brilliantly succinct statement of the aim of the experiment without making any direct allusion to its importance:

> The fundamental objective of the relativity mission is to measure the angular rate between the local frame (free-falling about the earth) and distant inertial space (defined by the "fixed" stars) to an accuracy better than 0.5 marc-sec/year, independently in each direction, for a one year experiment.
>
> (Heifetz et al., 2000, WeC11)

As can be seen in Figure 3.4, it was decided that for the actual experiment, the SIA would contain 4 gyroscopes, not just one. They were to consist of 3.81-cm-diameter homogeneous[43] fused-quartz spheres coated with a uniform[44] 1.27 μm layer of niobium.[45] This built-in redundancy[46] was a way to overcome the possible disaster that failure of just one gyroscope would represent for the experiment, had it been the only one. It also potentially allowed the team to refine their overall final results by combining the (4, assuming all the gyros worked correctly) different datasets. Each gyroscope was thus to measure both the geodetic effect and the frame-dragging effect independently of the other three gyroscopes.

[43] The impurities in the fused quartz used to make the gyroscope rotors are less than 2 parts per million. This ensured that stray forces within the rotors due to inhomogeneity were kept to an absolute minimum.

[44] It is precisely a flaw in this uniformity that was eventually detected as the source of the errors and excessive noise that the initial "raw" GP-B data contained

[45] For a brief summary of the characteristics of the gyroscope rotors see Buchanan et al., 2000.

[46] This is just one example of the inbuilt redundancy that permeated every level of the on-board systems and the entire experiment. For example, all on-board computing was duplicated so that one computer was always ready in standby mode in case there was a malfunction with the computer that was in use at any moment, which could then be rebooted and reprogrammed if necessary, from mission control, once the back-up computer had taken over control. The team actually did need to switch to the "B side" computer during the Science Phase of the mission, and soon afterwards were unlucky enough to have to reboot again. This resulted in the longest loss of science data during the whole experiment, as illustrated and explained in Figure 4.1.

GP-B was finally launched successfully from Vandenberg Air Force Base in California on 20th April 2004. The satellite was released into its polar orbit at a height of 640 km so accurately that it was not necessary to use the thrusters that it was equipped with in order to achieve the required orbit. However, problems with 2 of the 16 thrusters on the satellite and difficulties in locking onto the guide star meant that the IOC phase[47] of the mission had to be extended. This phase was originally planned to last 8 weeks but in the event it required more than twice that: a total of 129 days. This meant that the remaining liquid helium on board would not be enough to complete the initially planned 16 months of the Science Phase, which was reduced to slightly less than 12 months. Despite this reduction in the Science Phase of the experiment, the project team claimed that the data collected over nearly a year (it turned out to be 352 days) would be enough to produce results to the desired accuracy for both of the effects.

Reasons for doubting the experimental possibilities were expressed very clearly as early as 1986 by Van Patten and his team in their exposition of the state of data processing simulations. At a time (with VLBI in its very early days) when the guide star being considered was Rigel, they say:

> Even with such a "perfect" gyro, however, the question arises: Can the relativistic drifts be detected in the presence of random measurement noise, and other error sources such as satellite attitude control system errors, Rigel proper motion uncertainty, drifts due to gyro suspension forces, drift of electronic parameters such as instrument scale factors due to the thermal effects?
>
> (Van Patten, DiEsposti, and Breakwell, 1986, p. 157)

Although such doubts and fears were allayed and the mission went ahead, in hindsight, considering the problems the mission actually encountered, they seem to be almost prophetic.

[47] Although this was the term used throughout most of the history of the mission and during its actual execution (although there were some variations as it is sometimes called the "initialization and orbit checkout phase" (Li et al., 2015, p. 14)) in their 2015 review of the project, Everitt et al. (Everitt et al., 2015, p. 5) refer to this simply as the "set-up" phase. However, since the original "IOC phase" is far more descriptive and useful than calling it just the "set-up phase," I stick to the original term.

3.5 Expectations and Possible Interpretations of the Results

There were several foreseeable possibilities for the actual GP-B results when they were published, which on launch was expected to be in 2007. The results could have agreed, to within the limits of precision of the experiment, with the values predicted by GTR. Alternatively GP-B could have failed to detect any effect; or it could have detected an effect that fell outside the range of values that would have agreed with the PPN predictions of GTR as set out in Chapter 2. Since in the GP-B set-up the two effects were at right angles to each other, they were to be detected independently of each other, leading to 9 possible combinations for the results. Before moving on to consider the actual results in Chapter 5, there are a few general points that I want to make concerning the "expected" results.

I should first mention that in one sense the experiment looked, when data processing started, independently of the initial results that may have been gleaned during the space mission, set to be a failure. It appeared that it would not be the most accurate measurement to date of the $|\gamma - 1|$ PPN parameter; neither would it be the first measurement of the frame-dragging effect! These two cases were historical accidents and could not in any way be blamed on the experiment itself. Measurements of the Shapiro time delay claimed to have set a limit on $|\gamma - 1|$ that was beyond the expectations of GP-B (Will, 2014). Furthermore, the frame-dragging effect had apparently been measured. So, although GP-B could still hope to increase the accuracy to which the effect had been measured, it was difficult to see it as the first such measurement. Nonetheless, as I indicate above, it was still set to be the first experimental measurement of the effect: not just resulting from observation of a system beyond our control. There is an important distinction that is often made between merely observing and actually manipulating systems so that they react according to our design.

In the event, the complex levels of modelling involved led to a qualitative reduction of the levels of confidence in the result due to the addition of a whole new layer of analysis when it was realised that the data just were not

as expected. Even long before that revelation, when talking of the pre-launch stages of GP-B, Everitt explained it this way:

> Each result has been analyzed with unusual care, but some reserve is appropriate since each depends on elaborate data modeling.
>
> (Fairbank et al., 1988, p. 598)

Just as this is true when we consider the process of constructing and manipulating the models that allow us to study these phenomena, so it is true with the involved mathematical calculations that took place both in the processing of the data and the calculations to determine the other effects that interfered with the target signals in the actual results. With every parameter that is introduced into the calculations, a potential source of error is introduced. I think it is worth quoting Morrison and Morgan at some length on this point, when they discuss how the combination of actual apparatus and representational models measure the acceleration due to gravitation, G, as the parallel with GP-B is strikingly clear.

> It is possible using a plane pendulum to measure local gravitational acceleration to four significant figures of accuracy. This is done by beginning with an idealised pendulum model and adding corrections for the different forces acting on different parts of the real pendulum. Once all the corrections have been added, the pendulum model has become a reasonably good approximation to the real system. And although the sophistication of the apparatus (the pendulum itself) is what determines the precision of the measurement it is the analysis and addition of all the correcting factors necessary for the model that determines the accuracy of the measurement of the gravitational acceleration. What this means is that the model functions as the source for the numerical calculation of G; hence, although we use the real pendulum to perform the measurement, that process is only possible given the corrections performed on the model. In that sense the model functions as the instrument that in turn enables us to use the pendulum to measure G.
>
> (Morgan and Morrison, 1999, p. 22)

The case of GP-B is very similar; but in that case, a series of very complicated models were constructed to study the curvature of spacetime around the

Earth and its interaction with a spinning body. No matter how precise the experimental apparatus is, the accuracy of the results will always depend on correct analysis. Once it was discovered that the data from GP-B were not as expected and involved massive amounts of noise, a whole new strategy was developed to remove the noise and leave a "clean" or cleaned-up signal containing the trace of the two target effects. The additional problem that the team then had to face was that it appeared that they were, or at least might be, using the expected results to decide what to remove from the noisy data and thereby to arrive at what could be called "overfitted" (to put it mildly) results that were of no worth. I will consider these issues in Chapter 5 and Chapter 6, but here I just wish to consider a few general issues.

One such issue concerns the relative virtues of prediction (or novel results), such as the bending of light around a massive body predicted by Einstein, to those of accommodation (or non-novel results), such as GTR's accounting for the perihelion advance of Mercury. This is an issue that I treat in some detail in the next chapter. For now, suffice it to say that in the discussion regarding which of the two is more desirable in a theory, the ideal situation appears to be one of balance: a good theory should do both, but neither to the exclusion of the other. One of the crucial considerations is how many parameters are available in the model for the scientist to vary or tweak. If there are too few such coefficients or unknowns for us to adjust, then we possibly have a rigid, excessively linear theory or model that may be too strictly bound to the known instances it was designed to cover and may not have the flexibility necessary to provide a good fit between the model and novel data actually collected once it has been designed. If, on the other hand, we have too many parameters that we can vary, we will be able to adjust our model, or theory, to fit almost any data and it will be of little use in making predictions. This latter case is the worry that arises with GP-B as there were so many parameters involved that whatever the data, it may have seemed that the team could adjust the analysis of the results and the calculations and make them fit whatever conclusion we wished! To put this another way, they could have continued to tweak their models until they arrived at "expected" results.

As an addition to this, when an excessively large number of different factors

comes into play, the beauty of simple theories is diminished. GTR is indeed a simple theory (at least in its axiomatic expression); but the design of extremely involved experimental set-ups to isolate tiny local (weak-field) effects from the myriad of other factors affecting the apparatus, results in a loss of beauty. Even in what is, on the face of it, a quite straightforward and direct experiment, the introduction of so many additional parameters and new models—either in the chain linking "raw" (level 1) data to the final result, $\delta\theta$; or against which the experiment had to be calibrated—inevitably affects the final result.

The data that were collected during GP-B will, of course, still be available and open to further analysis and future reinterpretation. Regardless of the actual conclusions the project team published, it seems almost inevitable that the data will be re-examined in the future. Interpretations may well change and possible different effects or inaccuracies that were not taken into account may be added or removed in the future. Throughout the history of science, and particularly experimental gravity, the same results have been used at different times as proof (or at least evidence) of differing views.[48] Given this situation, I think that it is almost inevitable that the data will be reappraised in the future; the gravitational results of the experiment, as published by the project team, may well not be definitive. Any one small change in the numerous parameters in our modelling of the response of any of the systems involved in arriving at the results could lead to a change in the final outcome. Likewise, a change in our understanding of the underlying theory (or theories) could also give rise to a reinterpretation of the results.

On a different note, the fact that the movement of the guide star was tracked independently was heralded as a check on the independence of the scientific results. The two sets of data were originally to be maintained completely separate until each of the teams had completed all their calculations. The

[48]For example, in 1993, controversy arose concerning the 1919 Eddington eclipse exposition. At the time of the original expedition, in the wake of WWI, it has been suggested that one motivation for the the mission being hailed as such a success was the desire to heal wounds between Britain and Germany through a British expedition providing the necessary evidence to prove a theory proposed by a German scientist. In their book, Collins and Pinch, 1993 make a very strong case for the argument that in fact the results of that 1919 expedition were far from conclusive; and they reinterpret them as actually providing no support for GTR at all.

final experimental results for the drift of the spin axis could not therefore be known until the two datasets were combined (presumably this was to be performed by some "impartial" NASA representative). In the end, due to the changes introduced into the data processing this was not actually the case; but the intention was there initially and deserves comment. On the face of it, such practice may have seemed to guarantee the impartiality of the results; but I think it tells us much regarding the concerns of those who were in charge of the project or overseeing it. The measure certainly seemed to be designed more to guard against sceptical criticism claiming that the multitude of parameters involved in the final results had in some way been manipulated to give the desired results. (Whatever "desired" results might be in a situation such as this.)

This tactic can be compared with the use of double-blind testing in medical trials, for example. In double-blind trials, not only do the patients not know whether they are taking the drug being tested or the alternative (whether that is a placebo or the standard treatment), but neither does the supervising doctor. The technique was devised to stop doctors subconsciously looking for change in those patients who they knew to be in the treatment group, as opposed to those who had received a placebo; or to avoid any change in doctors' attitudes towards certain patients. It ensures, as far as possible, that no bias enters the trial from these possible causes. In the case of natural science, the subject of our experiments is always "blind"! It is an extraordinary step indeed to ensure that the experimenter too is "blind" in this sense. It certainly does not seem to be a measure designed to improve the scientific results. Are we supposed to believe that if the experimenters knew beforehand how the guide star had apparently moved, it might have influenced the patterns that they saw in the data or their interpretations of the data?

Of course it was understandable for some people to want to see this supposedly transparent check on the objectivity of the results in place to combat any claims of fixing the results. However, I see it as a very telling sign of the low prestige of science and scientists in general within society at the turn of the 21st century, and the defensive attitude adopted even by powerful scientific institutions with regard to their work. Furthermore, it must be said that within

a small, closed community such as that involved in experimental astrophysics in the US, it was hard to believe that nobody on either team would have any knowledge of the ongoing work, and partial results, of the other team until final completion. From a more cynical point of view, this additional degree of control on the publishing of the results could be seen as allowing whoever is in charge of combining the datasets and arriving at a final result to decide on the precise date of their release. This would then have allowed them to maximise publicity and maybe also maximise the effect they had on the next US federal budget, for example. However, with the timescale of this project, the recurrent delays, and the care that NASA usually takes over releasing its results, this hardly seems a relevant consideration in this case.

To close this chapter, as a prelude to the following more overtly philosophical chapter, and before moving on to consider the actual GP-B results and the amazing feat that the data processing actually represented, let me consider for a moment how we might regard the possible GP-B results using a Quinean-type web of knowledge structure as an analytical framework.[49] In such an approach, the knowledge that we produce is all interrelated and mutually supported: the closer one piece of knowledge is to other pieces of knowledge the more it relies on them. If we consider our web to spread out from a solid centre where we locate our most fundamental, unchanging knowledge, logic and mathematics, then this central area acts as a steadfast support for all the other knowledge stretching away from it. As we move out from the centre, so our knowledge becomes gradually less secure. We move through the findings of our well-established fundamental theories which support each other and are supported by (and in turn add support to) the central structure. They in turn form a (slightly less solid) base for the newer additions that in turn reach further out from them in different directions, representing the different fields of our knowledge (that seem so isolated at times, as I mention in the Introduction to Chapter 2).

As we continue to move out we come to still newer findings which are supported by many links to other theories but each of these is now becoming more

[49] I return to this analysis in Section 6.2, where I treat it in more detail in the light of the material in the intervening chapters.

tenuous as we are further removed from the solid central support. Finally as we approach the edge of our web of knowledge, we encounter the peripheral, uncertain, newest areas of knowledge and representations of nature. Here are the tentative findings of theories that are still trying to establish firm ground on which to stand and build support from all the other knowledge that is around them.

Now the question is: where can we see (or should we place) the knowledge provided by GP-B? I do not think that there can be much doubt that it is floundering on the outskirts of our web. The findings of our fundamental physics theories—quantum physics and GTR—are well established within the realms to which they apply. We are aware of the incomplete and transient nature of those theories, but they certainly provide a firm base on which to build. Despite being a straightforward test of GTR, with its respected position within our knowledge structure, the knowledge we gained from GP-B rests on new technologies and analytical models with few connections—as yet—to other knowledge. It has dozens of tentative connections to knowledge which is itself still in outlying regions and it remains to be seen whether these connections will strengthen with time and use, and the GP-B results will become a firm part of our knowledge structure; or whether alternatively one or more of the connections will prove to be a weak link that shakes the results of GP-B, or just what consequences that would have. With the new data processing techniques that the team were forced to develop to overcome the severe restrictions imposed on them by the nature of the actual data they collected, the GP-B result has certainly been drawn further from the central firm structure of our web of knowledge and indeed it may appear that whatever other connection there may be, everything hangs by the thread that represents these new, otherwise untested, data processing techniques and theories.

Of course this is not necessarily a bad thing—for human knowledge; it certainly is for GP-B—and is to a degree inevitably how we build up new knowledge. Slowly, as we become more confident in the work on which experimental results rest (through increased use and agreement with prediction, or adjustment to fit other new knowledge) so they become more firmly established as

part of the structure against which to test new hypotheses. Inevitably, experiments such as GP-B which are conceived as truly historic, groundbreaking steps must take their place within our broader framework of knowledge. The results do not stand alone, but rely on an elaborate interdependent system of observation, interpretation, modelling and theory.

From a distance, with a historical perspective, it may well be the case that we mistakenly attribute too much importance to an individual experiment. On further investigation and reflection, such crucial experiments invariably turn out to form just one part of a larger picture which includes many factors leading to an important change. This, for example, is often considered to be the case of the Michelson-Morley experiment. It is cited as the groundbreaking experiment that marks the end of the aether model and beginning of the road to relativity. On further studying the case, it can be argued that in fact it was at the time far from the decisive step so many have claimed it to be. In fact, it was considered by Michelson to have been a failure and is not believed to have been particularly influential in the genesis of STR since Einstein appears not to have been explicitly aware of the results until after 1905. At the time, it was very uncertain and on the outskirts of our knowledge structure. It was only as connections were made with other knowledge that it gained the support necessary to be considered a historic turning-point.

This may well be the case with GP-B. We simply do not know at the moment. What type of reception the GP-B results will receive from future generations we just cannot tell. They may gradually gain support from increased confidence in the work on which they rest. Alternatively, the results may be thoroughly revised in the coming years due to our changing knowledge base and the confidence we deposit in new and different findings. This is the way to build a solid knowledge base and whatever happens, we should make sure that we are prepared and willing to constantly re-examine previous results as necessary, and also to have the confidence in them that they deserve.

In the next chapter I consider in greater detail the philosophical framework within which GP-B was born, as well as the prevailing philosophical perspectives throughout the lifetime of GTR. That will then allow me to analyse the actual GP-B results in detail in Chapter 5 and Chapter 6.

Chapter 4

Counterfactual Difference Makers and Severe Tests

4.1 Introduction: Whence We Came

The century that has passed since Einstein published his GTR has seen great changes in our approaches to the analysis of science. Those changes have been driven in part by the scientific theories we have developed themselves and of course also by the direction that the advance of science (and its growing entanglement with technology) has followed. As a fundamental theory of physics (and in reductionist terms one could therefore argue that ultimately, as a theory of everything), GTR can be seen as playing an important role in those changes; especially in its initial years. In the early 20th century, both relativity and quantum physics introduced earth-shattering changes into the way we view the universe. They relied upon the novel practical application of mathematical methods that had been developed as purely theoretical constructs in the nineteenth century. The resultant revolution in our entire scientific outlook, from almost naked-eye observation and the application of geometrical methods to the postulation of curved spacetime and non-locality, leads me to consider this as a crucial point of departure for current physics, science and technology in general, and also philosophy of science. In this chapter I introduce some of the most important concepts and ideas that I will use to analyse GP-B in later chapters. But before I explain those, I want to recap some aspects of the road that we have followed

over these 100 years of living with GTR, which I believe help to explain and highlight the advantages of the stance that I will adopt.

As is well known, early in the 20th century, the syntactic approach[1] to the analysis of scientific theories was developed (also known as the sentential or law-statement approach). It formed an integral part of the "Received View" (RV) of science, endorsed by logical positivism[2] during the 1920s and 1930s and adopted wholesale by Western logical empiricism in the post-WWII era.[3] The history and development of the syntactic view of science has received much academic attention, as have the problems associated with it and its fall from grace, so to speak. Indeed, since Michael Friedman's 1999 book (*Reconsidering Logical Positivism*) we could say that we have even progressed far enough and acquired sufficient academic maturity to re-assess its importance and consider whether we have not moved too far away from some of its basic ideas. This law-statement conception can be seen as particularly relevant to GTR as it was developed when Einstein's theory was novel for scientists and philosophers alike, and partly as a response to it. It is clear that Einstein's work was not just highly influential but foundational in the thought and theories concerning space and time—and therefore the status of *a priori* and empirical knowledge—of leading logical empiricists including Moritz Schlick

[1] Both the syntactic and the semantic "approaches" are also known as the corresponding "conceptions" and "views"; there is no change of meaning intended or denoted by these alternative terms and I use them indiscriminately throughout the text.

[2] Although Stadler (1998) makes it clear that the term "logical positivism" did not appear until 1931 in an article in the *Journal of Philosophy* by Blumberg and Feigl, I use it anachronistically to refer to the philosophy of the Vienna Circle (a term which itself was not published until 1929) and the discussion group organised by Moritz Schlick starting back in 1924.

[3] The two terms "logical positivism" and "logical empiricism" are often used indistinctly. In this paragraph I use the former to refer to the early, European stage of the movement before the outbreak of WWII, and the latter to distinguish the period after the collapse of the Vienna Circle when migration to the Anglo-Saxon world of many of the Circle's members led to its philosophy becoming a major influence throughout the entire Western world after WWII. In the rest of this work, I simply refer to logical empiricism (and logical empiricists) to encompass features common to both periods.

and Rudolf Carnap.[4] Indeed, Einstein himself has often been labelled a logical empiricist. It should therefore come as no surprise that aspects of GTR can be seen as fitting in with such a syntactic approach in a way that is not at all typical of the majority of science, before or after this period. Despite providing the backdrop to and much of the reasoning behind what was to follow, I do not use this approach to any great extent in my later analysis of GP-B; so I just want to comment briefly here on how and why we moved on from the position it represented to the evolved scenario I situate myself within.

According to the syntactic view logical empiricists adopted, a theory is a system of axioms (or premises) together with indications of how to relate those axioms to totally separate empirical observations, and vice versa. In their efforts to eliminate what they considered to be unverifiable statements and meaningless metaphysical concepts from scientific theories, logical empiricists considered theories to be, above all, linguistic constructs. As Giere—one of the fiercest critics of this view of theories—puts it both extremely succinctly and generally when describing this approach:

> There were two components to the logical empiricists' picture of theories: a purely formal calculus, and 'correspondence rules' that link terms in the formal calculus with terms antecedently understood.
>
> (Giere, 1988, p. 74)

In the division of theories into two discrete parts that Giere describes here, the axioms contain the formal calculus and include the mathematical laws of physics, and often the initial conditions necessary to define a system. They are expressed using mathematics, symbolic logic and other elements of a purely theoretical language which contains no observational terms (at least in the strongest expression of this approach; though precisely one of the most widely recognised problems with it consists of identifying exactly when a term becomes observational). It is thus a characteristic of this approach that to fully express a theory we require two mutually exclusive lexical sets or two separate languages: a set of theoretical terms to form the purely theoretical

[4] Friedman, 1999, shows the importance of relativity to the development of (what he terms) logical positivism in his reappraisal of that stance at a time when, as I say, it had lost its former popularity, at least among philosophers of science.

language, and a set of observational terms. The methods of calculation and reasoning which the theory adopts likewise refer distinctly to observational and theoretical calculi: the way theoretical objects are dealt with (normally considered by the logical empiricists to be instruments to aid prediction and understanding, but considered not to represent anything real beyond the confluence of empirical observations) is different from the treatment of the empirical objects. The very nature of the correspondence rules (or C-rules)[5] that Giere names in the quote above means that they are necessarily expressed using both theoretical and observational language. Their purpose is to provide the link between observation and theory and thus to (partially) interpret the theory and bring "genuine" meaning to the theoretical terms.

It should be clear that this analysis—as well as the hypothetico-deductive (HD) method that stems from it—is simply not a description of how science works; and neither is it supposed to be. Rather than a naturalistic[6] analysis of theory construction and testing, it is reconstructive and quite deliberately starts with an uninterpreted formal calculus and purely theoretical terms to which (partial) interpretation is then added. Emphasis is also placed on deductive logic: given the axioms and the interpretation provided by the C-rules, we should be able to arrive at logical, deductive, conclusions. In actual fact, science usually starts with observation of regularities in nature

[5]Quite a range of terminology is used to denote this part of a theory; it is known as "correspondence rules," "correspondence laws," "bridge principles," "bridge laws," or simply "the dictionary" or "dictionary terms" as it is sometimes considered as a means of translating backwards and forwards between theoretical and observational languages. It is also sometimes called "coordinative definitions," although there is some discussion concerning the identification of this term (as used particularly by Reichenbach, (Reichenbach, 1958, chapter 1; section 4)) and the C-rules that we encounter in the RV.

[6] I use the term "naturalistic" throughout this work to denote practices that are based on and attempt to account for actual scientific practice, as opposed to efforts to reconstruct some idealised prescription of what science should be and how it ought to work. It is not meant here to suggest the psychologism and abandonment of foundationalism that Quine and others introduced into their programmes when talking of naturalizing epistemology. Rather I hope to capture the essence of later moves to include in our analysis of science anthropological and sociological points of view, instead of abstracting away from the actual process of the production of scientific knowledge and considering idealised logics of discovery, evidence, belief and knowledge.

and uses ampliative inductive logic (and imagination) to attempt to describe underlying, natural causes or propensities (what I prefer to see as phenomena) and suggest laws behind them. While this type of logical deductive analysis may be a useful check on the consistency of theories, it is clearly inadequate as a description of how the scientific process progresses. Furthermore, just how observational consequences can be deduced from theoretical language, even with the intervening C-rules, is anything but clear.

As an analytic tool, this syntactic approach has proved very useful, particularly in making explicit the assumptions that are contained in theories and how dependent they are on language. However, it has serious limitations. As I have indicted, some of the most important criticisms of this law-statement approach include precisely the clear division that it requires between purely theoretical terms and purely observational terms. Moreover, the vague idea it contains of partial interpretations seems to have eluded all efforts at clarification. These problems, together with those that arise from the verificationist ideas which formed a central part of the logical empiricist programme led to the rejection of the RV as insufficient by the majority of analysts, and the rise in importance of the semantic approach in the 1960s and 1970s, which emerged in direct opposition to the syntactic view and attempted to right some of its wrongs.

Thus, the primary aim of the semantic approach (also known as the structuralist or model-theoretic approach) was to overcome the perceived problems of the RV. The main problem is how to form a link between the theoretical terms laid down in the axioms of a theory and the actual natural system the theory attempts to represent, thereby providing theoretical terms with genuine meaning. For example, Suppe tells us that the semantic view of theories

> ...construes theories as what their formulations refer to when the formulations are given a (formal) *semantic* interpretation. Thus 'semantic' is used here in the sense of formal semantics or model theory in mathematical logic.
>
> (Suppe, 1989, p. 4; italics in the original)

The idea that Suppe is expressing here is that in contrast to the syntactic approach, the semantic conception considers the meaning of a theory to be expressed purely through its content. That content is the collection of

theoretical models that meet the conditions of the theory and thereby are the expression of its meaning. Thus, advocates of this model-theoretic approach claim to see a theory not as a linguistic construct (as logical empiricists did) but rather as the sum total of the possible models that the theory encompasses.

In some ways this approach immediately equates to our intuition concerning what a theory is. By adopting a semantic approach we have moved away from the rigid definition of a theory as a specific linguistic construct and arrived at the idea of a theory as something that aims to capture the meaning contained in all the slightly different expressions that it may have. It is easy to see parallels here with the different solutions of EFE which provide the different metrics that physicists use to tackle specific problems when dealing with the effects and consequences of gravitation. Although each solution is different, they are still seen as part of one unified theory: GTR. Thus, although GTR can be seen as forming an important part of the scientific landscape logical empiricists were trying to capture and represent through their RV (and even as partly responsible for shaping the syntactic view), and despite the semantic approach growing up in opposition to that view, we can see that this model-theoretic approach also reflects and offers insight into the nature of GTR. However, capturing the plural nature of theories and the different expressions they can have certainly does not indicate that we have found an unproblematic means of explaining what a theory is.

According to this conception, a theory does not seem to be a well-defined entity. Adopting the syntactic approach, a theory was deliberately defined as the linguistic construct resulting from the conjunction of its premises and correspondence laws. It was part of that conception of theories that a complete theory could (potentially, at least) be stated in full (although new empirical opportunities always allow for extension). This may not be very true to life in the sense that when scientists talk of a specific theory they usually do not have a well-defined, clear-cut structure in mind; but the sentential conception does at least contemplate an exhaustive statement of a theory. The semantic approach, on the other hand, either leaves us with a (set-theoretic or state-space) definition of a theory simply as the sum total of its theoretical models without making it clear what those models share in common, or it requires us

to define the boundaries in an axiomatic fashion; but quite how those axioms could themselves form part of the models they delimit as belonging to the realm of the theory without originating outside those models, once again, is anything but clear. Thus, this view leaves us with an idea of a theory not as well-defined, but rather as a population of models which can ebb and flow in different directions depending on exactly what we require of it. While this may appear to be a beneficial characteristic is some ways, it does not conform at all to the conception that we have of a theory, based on what may be considered exemplars, so at the very least it has problems as a general description.

While these two different approaches have taught us much of how science works and how we can analyse it usefully, both approaches are incomplete. Despite this, in certain circumstances one conception may prove more useful than the other; and they can both be efficacious means of analysing certain aspects of theory, and working towards an understanding of how theories contain meaning. As I have said, it may seem paradoxical that GTR, which exerted an influence on the logical positivists, lends itself very well to analysis as a collection of models; but it does clearly require the addition of the fundamental syntactic content of the relationships expressed in EFE. The task at hand, however, is the analysis of GP-B, a real experiment, together with the results it produced and their treatment and interpretation. As an experiment in fundamental physics, it is intimately bound to the problems of bridging the gap between theory and observation. As an attempt to confirm predictions of GTR, it could be seen as ideal territory for the adoption of a syntactic approach. As an example of analysis using the PPN formalism, it may seem that the best way to analyse the episode would be by considering a model-theoretic view. However, I will move away from both these rigid positions and instead adopt a more inclusive and naturalistic approach, as I have mentioned above, of analysing it first and foremost as an actual experiment and considering how the working scientists on the ground addressed the problems that arose and sought solutions to them. Taking on board the lessons of the previous approaches but moving on from where they left us, I therefore adopt the framework of New Experimentalism, as Ackermann

named the conception I refer to in the rest of this chapter.[7] Therefore, before I examine the results and implications of GP-B in the following chapters, I will summarise in this chapter what I consider to be the three most important aspects in this approach towards the analysis of scientific practice and present the general scheme of how I see them coming together to allow analysis of how knowledge is actually generated and gained through experimentation, at least in the case of GP-B.

The new direction of this New Experimentalism is, I believe, partly captured in the approach articulated by James Bogen and James Woodward in 1988, and in later work by the two of them and particularly by Woodward alone. It is also present and presented in the work on error as fundamental to the development of novel knowledge through experimentation and the importance of stringent testing of hypotheses by Deborah Mayo (1996). Although they approach the field from different directions, I see these sources as being complementary when it comes to assessing actual laboratory practices and I will combine elements from both in order to develop my own analysis. In doing so, my aim is to follow the naturalistic lines that these authors advocate and thereby prepare the way for providing an in-depth understanding of the actual aims, methods and practices of the scientists involved in GP-B in the remaining chapters. Moreover, as an exemplary case of the intricate links between formal methodological aspects of theory and experimental practice, GP-B can be seen as supporting the claim that prevailing out-dated epistemological attitudes among scientists are insufficient as a basis for decision making and can prove to have a negative effect on the advancement of science.

Thus, in the next Section (4.2), I offer an exposition of my preferred approach to the relationship between data and phenomenon, as originally put forward by Bogen and Woodward, 1988, towards the end of the 20th century, and which Woodward, 2011, has recently reaffirmed and defended from criticism. This represented somewhat of a break with the past as I have sketched it here in this introductory section: though more of an evolution towards person-centred methods and perspectives than a revolution in our analytical

[7] Ackermann uses the term as the title for his review of Allan Franklin's 1986 book *The Neglect of Experiment* published in BJPS (Ackermann, 1989, p. 185) and this seems to be the first recorded use of the term.

approach. The idea is to analyse what scientists actually do and how they generate evidence in favour of theoretical hypotheses and claims. This work, together with that of certain contemporaries, changed the focus of the philosophical debate that had preceded it and poses some quite different questions. Then, in Section 4.3, I introduce the second aspect which Woodward went on to argue for, building—as he tells us—on the work of David Lewis and Robert Nozick: counterfactual sensitivity (Woodward, 2000). Woodward linked this concept to the reliability of findings and the reliabilist tradition in epistemology (particularly of Goldman, 1986 and Dretske, 1981) and demonstrated that it is the way in which experimenters track the truth of their claims concerning actual difference makers. In Section 4.4, I then present Deborah Mayo's work on the severity of the tests that hypotheses are subjected to through experimentation and the matching of data to those hypotheses (Mayo, 1996). Mayo defends non-Bayesian probability and what she calls standard error statistics, together with the role of both of these in the practice of science and the concomitant generation of knowledge. Mayo's work can also be seen as a naturalistic approach to the analysis of science. She considers the extremely varied and diverse toolbox of methods and applications that scientists use in the field to analyse their observations and justify their findings; and from that perspective she makes insightful observations and draws somewhat unorthodox conclusions concerning the advance of science.

The composite approach and interpretation I develop from these three ingredients provides a meaningful way of overcoming the perennial problem of bridging the gap between the abstract formal structure contained in the theoretical models we construct and express through mathematics, and which we use to represent our current scientific theories, on the one hand; and the data we collect to probe those theories through manipulating and observing actual physical systems, on the other. It forms the basis of much of my analysis of GP-B in later chapters.

4.2 Observation and Measurement: the Production of Data

I will now consider the first of the three ingredients that I require for the naturalistic approach (as I explain it in footnote 6 above) to the analysis of scientific practice and how it leads to the production of knowledge that I will use in my later considerations of the GP-B experiment. This initial constituent is an account of what it is that scientists actually observe and measure, and why. This may immediately seem both to be an issue that is far too general to be of much use in the analysis of a particular episode in science and to require an answer that is too specific to each case to offer an opportunity for any kind of general analysis. Of course, as we know and as is constantly emphasised from the perspective of New Experimentalism, the ways in which scientists proceed and the solutions that are found to resolve different issues that experimentation throws up are almost as many and varied as the scientists involved in these activities. So, what I will consider here is the specific angle or route that I have chosen as the most appropriate to analyse the findings and claims of GP-B. However, I also wish to emphasise here that this first ingredient represents a shift from the more common view that I believe goes a long way to demystifying how it is that experimentation and observation can support the theoretical constructs that lie at the heart of science in many and varied instances. Although it may be one of a range of analytical choices, it is one that I believe has a broad scope of applications and can beneficially be deployed in the analysis of an immense amount of experimentation.

This ingredient is the specific distinction between data and phenomena, and the roles they play in much of scientific practice, as propounded by Bogen and Woodward back in 1988,[8] in connection with the:

[8]The 1988 paper by Bogen and Woodard is where the ideas I relate here are first put together and published by those authors. However, it should in no way be seen as the only exposition of their views or indeed as being complete. Maybe most importantly in this respect, Woodward published "a restatement and defense" of the ideas in Woodward, 2011. Moreover, the original paper was itself just one of several that the two authors published together on the subject; and Woodward went on to work in more detail on different aspects of the ideas which I continue to borrow from in the next section (Section 4.3). I will comment at the end of this current section on

> disparate problems and procedures which are relevant to inferences from data to phenomena in various experimental contexts.
>
> (Bogen and Woodward, 1988, p. 314)

The crucial point as they state it here is that in many experimental contexts, inferences are made from data to phenomena, in direct contrast to the common HD model in which the relevant inferences in scientific practice are from theory, or theoretical phenomena, to data. There are several important aspects to this radical idea—rejecting the more standard view of science—and its consequences that require explanation and which I will go through one by one.

First of all, and central to their claim, the contention of Bogen and Woodward is that scientists do not always observe phenomena, as is often claimed, but what they certainly do during experimentation is record data. It is clearly the objective of all scientific enquiry to be able to treat and interpret the data so gathered in such a way as to use it to imply certain causes, propensities or phenomena; but the clarification of the distinction is crucial to their scheme and the way in which I will adapt and apply it. They emphasise that, as opposed to phenomena, when they talk of data they are talking about the actual readings and measurements taken by scientists in the laboratory or in the field. It is the data thus generated that they take to be the actual observable; on many occasions any phenomenon that may be behind the data and which scientists often say that they "observe" (meaning more literally "detect") is

---

what I see as being the only substantial qualification that needs to be made of Bogen and Woodward, 1988 in the light of Woodward, 2011. Here is what Woodward says of the initial exposition:

> Two decades ago, Jim Bogen and I published a paper (Bogen and Woodward, 1988) in which we introduced a distinction between data and phenomena and claimed that this had important implications for how we should understand the structure of scientific theories and the role of observation in science. This initial paper was followed by a series of papers on related themes (Bogen and Woodward, 1992; Bogen and Woodward, 2005; Woodward, 1989; Woodward, 2000
>
> (Woodward, 2011, p. 165).

considered by them not usually[9] to be directly observable, in the sense that it cannot be seen and it is not what is actually measured. While they make it clear that we certainly do *detect* (which seems to be what many scientists mean by "observe") theoretical entities, the theory that posits those entities in no way predicts what we actually *see* ("observe" in their more literal, everyday sense). Let me just quote them from both the start and end of their paper to make this clear.

> Phenomena are detected through the use of data, but in most cases are not observable in any interesting sense of that term. Examples of data include bubble chamber photographs, patterns of discharge in electronic particle detectors and records of reaction times and error rates in various psychological experiments. Examples of phenomena, for which the above data might provide evidence, include weak neutral currents, the decay of the proton, and chunking and recency effects in human memory
>
> (Bogen and Woodward, 1988, p. 306)

> . . . phenomena for the most part cannot be observed and cannot be reported by observational claims. In order to support the contention that phenomena are observed, terms like "observation" and "observation-sentence" must be used too vaguely to say anything informative about science.
>
> (Bogen and Woodward, 1988, p. 343)

The question of course arises as to why this distinction should be important. The point is that by claiming that scientists observe phenomena, in many cases we are missing out a vital stage in the scientific process. Thereby, we are failing to adopt the naturalistic approach that I mention above; but moreover, by reverting to an idealised reconstruction, we may overlook a step in the chain that leads from observation (perception) to conclusion and maybe knowledge. We cannot appreciate or assess the whole chain that the process consists of if we miss out vital links. So if we are interested in analysing the generation of knowledge through scientific practice, it is important to see where the

[9] As I mention at the end of this section, it is precisely this use of "usual" and other such qualifiers that Woodward feels was an overstatement and unnecessary for the point that he originally wanted to make. That point was an attempt to expand the analytical options; not to suggest that the framework presented here should be seen as the default norm; and certainly not as the only option. It is, however, the scheme that I adopt in my work here, so I stick to the original.

process starts and see what happens at the different stages. I hope that this will become clear when I return to it below (Section 4.3). For now let me concentrate on the contention of Bogen and Woodward that it is data and not usually phenomena that are observable (in the sense of "seen" or "perceived").

Making such a distinction between (observed) data and (typically unobservable) phenomena is in contrast to the way in which both scientists and philosophers of science tend to speak of and treat experimentation as observing the phenomena of interest (or attempting to). Bogen and Woodward, together with the approach of New Experimentalism in general, see it as essential to recognise and analyse the step that experimenters take from the production of certain data to claims of support for underlying theoretical phenomena. This is where the experimenter comes to the fore as an active agent taking decisions. In the laboratory, as well as all their recognised skills and acquired know-how, experimenters bring into play all their accrued tacit knowledge. The combination of what can be seen as flair or having a knack for experiments, much of which can probably be accounted for by tacit knowledge, together with the benefits of experience and of having worked with other dedicated and competent experimenters would seem to explain why some individual scientists shine among their peers and acquire well-deserved reputations as geniuses of the laboratory. It is precisely at this stage, that the elements that cannot be captured through logical relationships come into play. Experimental scientists are known for spending long hours working with, and often custom designing, their apparatus. This occurs on occasions long before any data are generated at all; or on others, in between data collection, in order to tune the apparatus and hone it to the job in hand. Just when the apparatus is working correctly and when valid data can be recorded is a decision made on the ground, in the laboratory; and it is at least in part here that the reliability of the data as tracking real effects is established.

The danger of considering that phenomena (not data) are observed is that this step in the process of the production of knowledge through experimentation, which consists of the actions of the experimenters, is glossed over. There are two steps here that may be conflated into one. According to New Experimentalism we have the step from data to phenomenon; and the step from

phenomenon to background theory. In contrast, the only move or step that is all too often considered is that from the (supposedly observed) phenomenon to the background theory. And this single step can be seen as conforming perfectly to the HD framework: phenomena often can be derived from the background theory; not so data. Here we can lose one of the vital links in the chain of knowledge production: that of the expert practitioner who brings more than logic to the laboratory. This is where a link is forged between personal understanding and appreciation of the experimental set-up, and the requirements of theory for the data collected to accurately track the target phenomenon. The practice of science is perhaps the most human of all activities; and nowhere do we see that more clearly than in the combination of both theoretical and practical knowledge, understanding and know-how that is necessarily brought together in experimentation to persuade nature to reveal intricate details of how the universe works through the reliable tracking of genuine difference makers. The counterfactual sensitivity that scientific advance requires is arrived at through intervening in the universe in extremely precise ways and knowing how and when to observe the results of that intervention.

If we gloss over this step, there are two equally misguided paths that we can stumble down with respect to inquiry into the functioning of science. On the one hand, analysis of claims concerning warrant and the reliability of scientific findings may focus on the internal mental aspects of the experimenters. In the case of much epistemology this is the path chosen, most typically concentrating on the psychological functioning of perception. This leads to questions concerning how scientists can know that they have observed the data, rather than considering questions of how they decided that their apparatus was working correctly and therefore when to start recording those data. While the philosophy of mind and surely medicine and psychology itself, as well as myriad other areas of knowledge can benefit enormously from this study, it seems clear that this is not where the key to the reliability of scientific knowledge resides. The other approach is that which may seem to be recommended by the HD approach and again adopted by much standard analytical epistemology: consider the practitioner as an ideal subject and concentrate on the logical relationships that propositions concerning phenomena stand in to the

axioms that form the nucleus of our background theory. Again, this study can be immensely fruitful, but it is missing the vital step that takes place between these two areas of research: between the internal perception of the scientist and the relation that the theoretical predictions regarding phenomena stand in to the more abstract background theory. This is the step that can be missed and is the one that connects the data to the phenomena.

Standard analytic epistemology then attempts to overcome some of the problems faced by a syntactic approach and span the gulf that separates observation and theory. It typically reflects on the logical relationships that hold between propositional content that is considered as evidence (principally derived from perception) and the propositions expressing theoretical knowledge that it is claimed are supported by that evidence. It is often thereby missing the point entirely that the strength of the evidential chain depends on the move the scientists have made from data to phenomenon; not on the logical connections of what observations can be derived from the theoretical propositions concerning the phenomenon in question; which is in many cases a question of logical deduction (and so not of interest to the study of the practice of experimentation).

Thus analytic epistemology aims to examine the status afforded to supposed evidence, normally stemming from perception, by way of the different logics of the evidential relationship (and also by the means through which it is generated, which I will return to in Section 4.3). Great weight is still placed on this type of analysis, sometimes called the evidential-relationship view (for example, this is how Mayo, 1996, refers to this approach) of how observation provides the grounds and justification for theory in science. The idea is sometimes expressed by claiming that the evidence that supports a theory, which clearly in all empirical science must have its origin in observation, is "entailed" by the theory; the logical characteristics of such entailment are then studied. This standard analytic epistemological approach does not seem to capture at all how experimental scientists typically (or at least, on many occasions) evaluate evidence.[10] Furthermore, the vital role of the experimenter

[10]Certainly not once a science has developed beyond the very early stages of finding general overarching conditions or regularities from which certain crude observations could indeed be said to follow. I am concerned with what are often

in the handling of apparatus and data is completely neglected. The actual human input, which must be so characteristic of science at all stages and which it seems contradictory to try to eradicate when it is human activity we are trying to analyse, is reduced to an absolute minimum and replaced with logical relations.

In practice, scientific evidence is not usually evaluated through logical relationships but through empirical discovery of actual difference makers (together with data analysis and interpretation that often involves complex statistics). It is through much manipulation and experience that scientists learn the skills and acquire the know-how necessary to determine when data indicate genuine effects. There are many judgement calls involved (and of course, we often get it wrong). This brings me to the second aspect of Bogen and Woodward's proposal: that data, what scientists observe and record, are not (generally; though there are clear exceptions) predicted in any specific detail by the theory those scientists are interested in; or indeed maybe not by any other theory or combination of theories. In their words:

> Data are, as we shall say, idiosyncratic to particular experimental contexts, and typically cannot occur outside of those contexts. Indeed, the factors involved in the production of data will often be so disparate and numerous, and the details of their interactions so complex, that it will not be possible to construct a theory that would allow us to predict their occurrence or trace in detail how they combine to produce particular items of data.
>
> (Bogen and Woodward, 1988, p. 317)

To understand the role of experimenters and just how knowledge can be generated through experiment, this aspect is crucial. Scientists do not (necessarily) design experiments to observe phenomenon (directly, in the sense I stipulate

considered "mature" sciences, which typically use complicated datasets that are usually the output of complex observational equipment together with methods that have been designed specifically for the purpose at hand, and are concerned with probing and expanding pre-existing theory via effects or phenomena that are not usually directly observable at all. (Since later I apply my analysis to the case of GP-B, this slight limitation will certainly be no drawback there.) As Bogen and Woodward (1988) remark: "... a scientific discipline which is marked by attempts to predict and explain what is observed is usually at a relatively naïve and primitive stage of development. (p. 307)

above) but to identify (or confirm or refute) genuine difference makers. It is only through the incredibly contrived circumstances of the experimental set-up that the target phenomenon will have a systematic effect on the data that is recorded and that can then be treated and analysed, usually using statistical means, to reveal that effect. One of the examples that Bogen and Woodward cite in this respect is (once again) that of the 1919 Eddington expedition to measure the bending of light as it passed close to the sun during an eclipse. The actual photographic plates that were produced by Eddington could in no way be predicted by GTR: they were fruit of a whole host of circumstances: the functioning of the equipment and the set-up itself.

Again, this is part of the break with the still more common view of scientific practice as advancing along a HD route and the evidential relationship as seen by standard analytical epistemology. According to Bogen and Woodward, data (the specific observations and measurements taken by scientists) should not be seen as what a scientific theory aims to predict. Rather, it is the (usually) unobservable *phenomena*, or certain characteristics of them, that have recurrent, reproducible effects on certain systems and that can therefore be predicted and, it is hoped, in some way gleaned from the data. These phenomena may well be derivable from a larger theory, but not so the data that are their actual observable manifestation. It is the phenomena that are (typically, maybe ideally; certainly in the cases considered by these authors) embedded within an overarching theory within which they stand in logical relationships to other parts of the theory. Thus, while the theoretically predicted phenomena remain unobservable, the specific data that will be observed cannot be logically deduced (or induced) from the target phenomenon or the wider theory. This understanding of phenomena is the third element of this view which I will adopt. The phenomena are what the data that is collected (ideally) support or refute: not what is observed. Again, it can be illustrated with the GTR prediction of the bending of light: the phenomenon that is a theoretical prediction of GTR that is behind the data collected by Eddington and it was hoped would be detected through appropriate handling of that data.

In this way, the data which are produced, often through the exceedingly

complex interaction of a target phenomenon and many other unquantified, possibly unexplained and even indeterminable effects, are regarded as possible evidence for an underlying unobservable target phenomenon. It is the job of the experimenter to design a set-up within which the target phenomenon will manifest itself in the data. Let me just repeat what I said in Chapter 3 (page 84) regarding the task of the experimenter in this respect. In any experimentation, depending on the options available and the convenience of tackling obstacles in one fashion or another, effects can be physically shielded against, sensitivity can be increased, or data handling can be refined to remove predicted signals and effects that are not the target of the experiment. We need to include the human practitioner if our account of the generation of knowledge is to capture how science has become the success it is.

So, the actual evidential support relationship between data and phenomena is empirical. Experiments are designed and observations are made in such a way that the dataset produced can be seen as evidence in favour of or against the target phenomenon having a causal effect on the system in question: being an actual difference maker. This was certainly the case with GP-B. It is the identification of these difference makers and the division of observation into signal and noise, or effect and artefact, that the experimental tradition within the philosophy of science attempts to analyse. When I examine the actual results of GP-B, I will be concerned with whether the team correctly identified the actual difference makers; and as I have already said, in the event they did not have experimental design entirely on their side when performing this task with the unexpected dataset they actually recorded.

This has led me into the second crucial ingredient that I need in my analysis: the identification of difference makers. This work seems to have been carried on mostly by Woodward alone, and it is from him that I will borrow what I present in the next section. Before finishing here with "Saving the phenomena" though, I just want to mention that, as I say above in footnote 8, Woodward revisited the work in 2011 to defend it from criticism, offer some clarification and make a few small rectifications. He stands by the vast majority of the work, but to be fair to him, I feel I must include the following rather long quote.

> For starters, we were far too willing to formulate our central contentions as claims about what "typically" or even always happens in "science" ... our goal was to provide a framework that made sense of various sorts of data-based reasoning, one that grants that such reasoning can be relatively independent of certain kinds of theory. In support of this framework, it is enough to show that reasoning from data to phenomena *can* (and not infrequently does) successfully proceed without reliance on theoretical explanation of data. ... What we should have said (I now think) is that phenomena need not be observable and that in many cases standard discussions of the role of observation in science shed little light on how phenomena are detected or on the considerations that make data to phenomena reasoning reliable.
>
> (Woodward, 2011, p. 171; italics in the original)

4.3 Difference Makers and Counterfactual Sensitivity

As I have indicated, the second ingredient for me must be the means by which genuine difference makers are identified through experimentation. It is no good theorising and collecting data if those data are not somehow tracking the truth, or falsity, of our hypotheses: science would never advance and knowledge would not be generated. As I reiterated in the previous section, the experimenter has a range of general choices available, always with the aim of generating and collecting data that will be sensitive to the presence (absence), or strength of the target phenomenon. Experimental design and set-up is so immensely varied that there is nothing that I can say as to the means employed; but the goal is always to reflect in some way the presence, absence or degree of some phenomenon of interest (or a specific effect of a phenomenon).[11]

[11] Of course, it would be virtually impossible for such a broad-sweeping catch-all statement to be strictly true! There is plenty of experimentation and whole classes of experiments that have no target phenomenon at all; rather they are purely exploratory in nature. I could try to generalise such experiments and bring them into my framework by claiming that the target they investigate is the existence of any effect at all within the data they generate; but I think it is safer just to add the somewhat

This requires the data produced to be dependent on the phenomenon, if it exists: for the phenomenon potentially to make a genuine difference to the data collected. Thus, an ideal experiment is one that indicates, through the data produced, whether the target phenomenon is present or not; or the degree to which it occurs. So, after executing our ideal experiment we can always say: had the target phenomenon not occurred, we would not have gathered the data we did. Or alternatively: had the target phenomenon occurred, we would not have gathered the data we did; or: had the target phenomenon occurred to a different degree, we would not have gathered the data we did. Whatever the objective of our enquiry, our ideal experiment, due to its perfect set-up and functioning, will always indicate through the specific data generated the presence, absence or degree of the phenomenon in this way. Of course, no experiment is ideal; but equally, all experiments (with the proviso in my previous footnote) aim at (and tend towards, if we are getting things right) the counterfactual sensitivity I have just outlined. It is this dependency of data on target that seems to me to unify the vast majority of experimentation, if not all. Moreover, and if we consider the immense spectrum of scientific experimentation, it seems that this may be the only thing that it would be possible to claim is universal to the aims of scientific experimentation.

This counterfactual relationship between the (unobservable) target phenomenon and the actual dataset produced through experimentation has been championed by Woodward, among many others. (As I have already said, I draw heavily here on Woodward, 2000, in which he notes particularly the prior work by Lewis (1986) and Nozick (1981), and also cites the work of Deborah Mayo that I will refer to frequently in the next section.). He relates this vital step in the chain leading to the generation of knowledge through experimentation to the reliability of the methods adopted: it is through genuine counterfactual

redundant proviso that I am interested here only in those experiments that do have a specific aim. The real irony of the situation, however, is that in certain respects which I will analyse in the following chapters, the case of GP-B can be seen as an experiment that ended up testing a hypothesis that was not its aim at all. As it turned out, here was an experiment which certainly did have an aim, but whose data were then used additionally to test totally different claims that had not been the primary objective of the data generation and collection.

dependency that experimentation produces reliable results. He presents the thesis as follows:

> The basic idea that I will be defending is that in many typical cases the relationship that must be present if data are to provide evidence for some phenomenon-claim is a certain sort of systematic pattern of counterfactual dependence or sensitivity of both the data and the conclusions investigators reach on the phenomena themselves.
>
> (Woodward, 2000, p. S166)

So, experiments are designed, and datasets are manipulated and interpreted in order to arrive at the necessary counterfactual sensitivity for the target claims (phenomena) to be supported or refuted (by the experimental data). Moreover, as I mention above, the most important tools that scientists use in their analysis of the actual experimental data produced, and the ways in which they determine whether or not those data support the target claims, are statistical. As both Woodward and Mayo correctly observe, it is largely through statistical manipulation that scientists actually arrive at their conclusions regarding support. In this way, as I point out below in Section 4.4, when Mayo talks of support (for hypotheses by evidence in the form of data) being a matter of degree, we can see it as being the degree to which the data statistically support the counterfactual reasoning that identifies actual difference makers. Woodward similarly tells us:

> As we shall see in more detail below, in the case of both measurement and detection, the extent to which this general pattern is satisfied will be a matter of degree—when or to the extent it is satisfied, I will say that the procedure and the associated evidential connection between data and phenomena are reliable.
>
> (Woodward, 2000, p. S167)

The idea of counterfactual sensitivity which leads to the reliability of experimental findings is the same as saying that the specific data produced are dependent on, or track, the truth or falsity of the target propositions. As Woodward again tells us:

> the detection or measurement procedure should be such that different sorts of data $D_1 \ldots D_m$ are produced in such a way that

> investigators can use such data to reliably track exactly which of the competing claims $P_1 \ldots P_n$ is true.
>
> (Woodward, 2000, p. S166)

Once again, this is a break with the idea of entailment offered by standard HD reasoning, which furthermore operates in the opposite direction. According to that view, the theory (or theoretical phenomenon as I prefer to see it) entails the data (or fails to), rather than the sensitivity of the data supporting (or not) the presence (or degree of) the phenomenon. This difference between counterfactual sensitivity and the more standard analytical approach can clearly be seen if we consider the following contrast. For evidence in favour of the hypothesis being investigated to be deemed to have been generated, both require that the specific dataset produced, D_i, stand in the appropriate relationship with the corresponding claim, P_i. However, whereas the logic of the evidential relation stops here, counterfactual sensitivity furthermore requires that were P_j (for every $P_j \neq P_i$) actually the case, then D_i would not have been produced. This difference is an attempt to tackle the perennial problem of the underdetermination of scientific theories. Let me illustrate what I mean with the following toy example.

The logic of entailment cannot distinguish between a theory, T, which stands in the appropriate relationship of entailment with the data D_i, and the theory T' which states that $T+$ *Christopher Evans is Welsh*. As is well known, in such cases, if the required relationship holds between T and D_i then it also holds between T' and D_i. In contrast, D_i fails to comply to the necessary counterfactual relationship with T (or T'): it is not the case that had T' actually been the case some dataset other than D_i would have been produced, while T' may be precisely the type of alternative that the experiment would be required to discriminate. To reiterate, according to the standard analytical reading, it is the mere occurrence of D_i that is evidence for P_i (or T in my toy example) irrespective of what other outcomes may possibly have been produced.

Another important point which the standard HD view misses is that the required counterfactual sensitivity is often an empirical matter for experimenters to determine by actual design and use of equipment. It is sensitivity to the underlying phenomena (or if one prefers, propensities, causes or laws)

and is often best established through altering experimental conditions in subtle ways. Although it is not (always) possible to determine what specific data it is necessary to produce, a substantial (statistically significant) change in the data produced may be observed by altering conditions and thereby affecting the phenomena that are actually making a difference to the data. (According to both counterfactual sensitivity and hypothetico-deductive reasoning, it is perfectly conceivable that identical datasets could be considered evidence in favour of some proposition in one instance, but not in another; because, for example, certain background assumptions may hold in one case but not in the other. However, it is only counterfactual sensitivity, and not logical entailment, that always requires in addition to the matching of data and phenomenon that had some alternative actually been the case, then the data that were produced would not have been). Thus, counterfactual sensitivity seems to be a far more useful way of understanding how experimental data provide evidence for phenomena than more traditional logical models.

I have argued for counterfactual sensitivity at some length as it is a feature that I will consider in my analysis of the complicated arguments employed by the GP-B team in the light of the anomalies they encountered in their data. Notwithstanding, by way of example it can be seen in action in the final claims they made regarding the support (or otherwise) that their data represent for the phenomena of frame dragging and the geodetic effect, and thereby for GTR. In this case it can equally be argued that this is a case of the more standard HD method in action: the phenomena (frame dragging and the geodetic effect) are detected through the experimental results being entailed by (agreeing with the predictions of) GTR. I still feel that this latter view is over-simplistic in that such an idealised reconstruction fails to come close to explaining how it is that the experimentation involved leads to the generation of knowledge. So, I offer the example here not as part of any argument in favour of the adoption of one or other approach, but simply as an illustration of how my preferred approach works and an indication of why I prefer it.

Figure 4.1 shows the north-south component of the final GP-B results. As I say, these can be seen as observation of the geodetic effect predicted by GTR; that is, we can say that this evidence is entailed by GTR (whatever it is that we mean

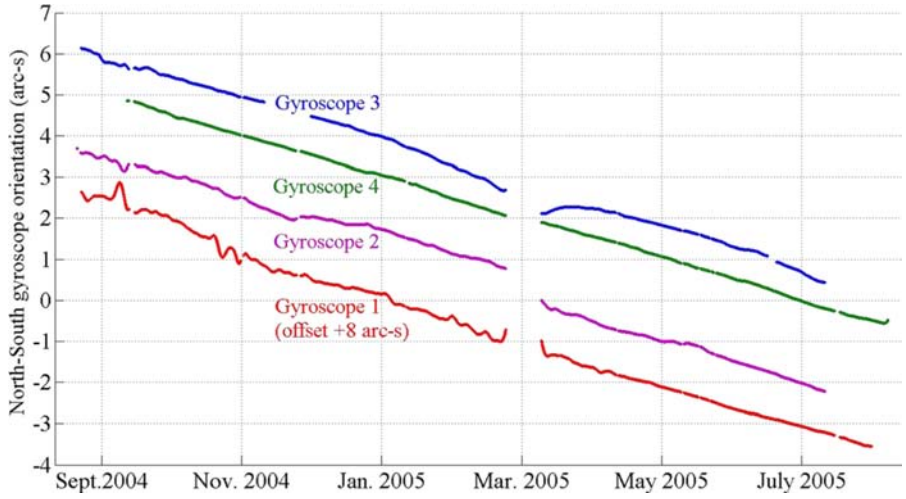

FIGURE 4.1: The evolution of each of the four GP-B gyros over the length of the year-long experiment in the north-south plane is shown. The data are divided into several separate segments for each gyro due to breaks in data collection either for the individual gyro or for the entire on-board data collection and recording system. The most notable is the break of over a week in March 2005. This was due to an initial multi-bit error (MBE) which triggered a switch-over to the B-side backup computer, followed by several MBEs within a few days causing a re-boot of the backup computer on 18th March. The results can clearly be seen to support the claim that each of the gyros drifted at a rate of approximately 6.5 arc-seconds per year constantly during the experiment, in agreement with the GTR prediction of the geodetic effect on the GP-B set-up.
[SOURCE: Everitt et al., 2015]

by "entailed by"). Moreover, as I explain in detail in the following section (Section 4.4), these data are used in a "novel" fashion here, insofar as they were not used in any way in the derivation of the phenomenon in question (there was no "double counting" of the data, as I explain below). Alternatively, we can adopt the following counterfactual reasoning to demonstrate the sensitivity of these data to the underlying phenomenon of interest:

CntFac If the GP-B gyros had not experienced a geodetic effect as predicted by GTR, then we would not observe the drift of approximately 6.5 arc-seconds in the north-south plane that is present in the year-long gyro

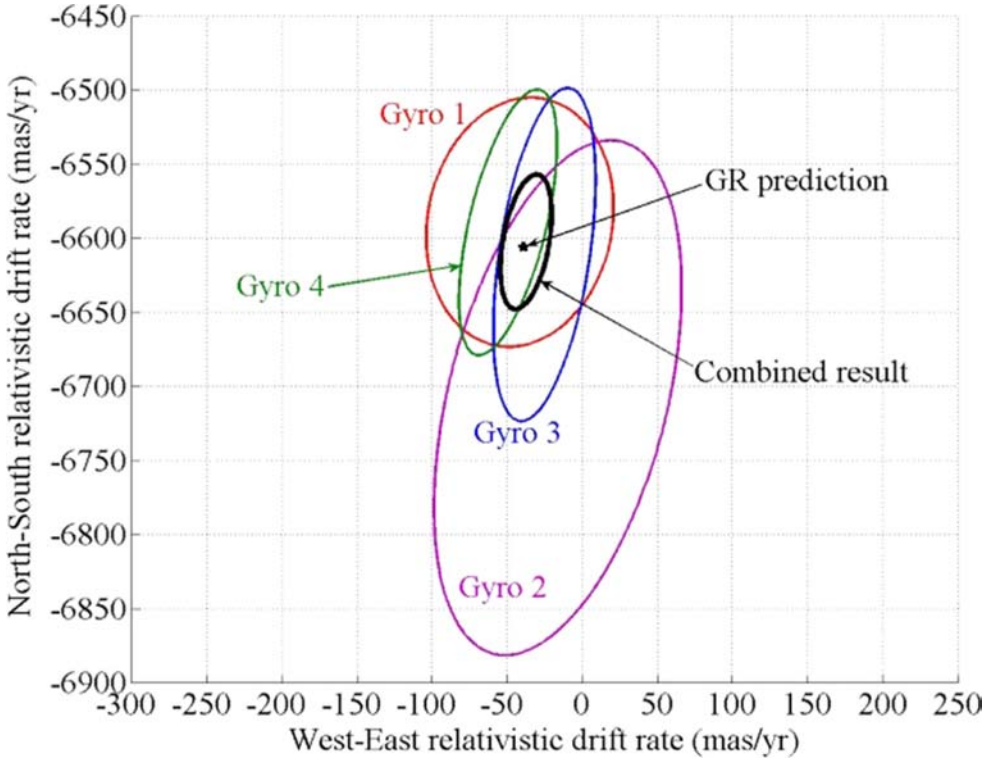

FIGURE 4.2: Overall GP-B results showing the individual results for each gyro and the overall combined result from all four gyros. The GTR prediction is also indicated by a star. [SOURCE: Conklin et al., 2015]

spin axis orientation data.

While both ways of considering this result are certainly informative, I believe that the latter, counterfactual approach is more complete. It tells us not just what did happen (the gyros drifted in the way predicted by GTR) but additionally that this would not have happened had GTR not held. In contrast the former, entailment conception remains silent on this point (maybe any metric theory of gravitation would have produced such a result: we just are not told). To see how this would work by degrees, consider Figure 4.2.

In this case (Fig. 4.2), we can see in a straightforward fashion that the GP-B results agree perfectly with the GTR prediction to within the degree indicated by the corresponding oval denoting the error on the data. (The situation is not nearly as straightforward as it appears when we simply consider these claims

without taking into account the tortuous path that was followed to arrive at them, as I do in the next chapter (Chapter 5). It is precisely there, when considering the actual sensitivity of the claims to the experimental conditions and not conflating it all as entailment, that the difference between the two approaches becomes interesting and highly informative). So again in this case a HD reconstruction could be as follows:

HD: To within the given error (of the order of 1% in the case of the geodetic effect and approximately 20% in the case of frame dragging) the observed drift of the four GP-B gyros over the year-long experiment is that entailed by GTR via the geodetic and frame-dragging effects that follow from that theory.

This is, of course, perfectly correct in as far as it goes (and allowing for some appropriate understanding of entailment); but it is not as informative, I maintain, as the corresponding counterfactual sensitivity statement:

CntFac: If the GP-B gyros had not experienced a geodetic effect as predicted by GTR, then we would not observe the drift of approximately 6.5 arcseconds in the north-south plane that is present in the year-long gyro spin axis orientation data.

Be that as it may, it is not until I incorporate the third, and closely related, ingredient of the approach I adopt, that this distinction becomes far more important. That is what I will now present in the following section.

4.4 Severe Tests

I have now introduced the first two ingredients of the approach I adopt in my analysis of the claims regarding the GP-B findings: the use of (experimentally contrived) observed data as evidence of a (theoretically predicted) target phenomenon or effect; and the requirement that those data be appropriately counterfactually sensitive to the target phenomena. So I will move on to consider the third, which can be seen as a variation of the second, specifically in order to rule out one of the criticisms that has been levelled at the GP-B

data analysis (see my discussion of the NASA, 2008 report in Section 5.1). The charge against GP-B is one of what has been called "double counting" or non-novel use of data. To broach this subject, I will start by considering once again the way in which a hypothesis we construct concerning the presence of (the effects of) a phenomenon can be said to undergo a test when we perform observations or experiments and collect data that represent evidence for (or against) the hypothesis in question. In such a case, the test consists of the match between the data we collect, and the data we would expect if the phenomenon was present, always being tracked via the appropriate sensitivity, that demonstrates the veracity of the hypothesis (the presence or effects of the target phenomenon). In this way, we can say that if the data match (in whatever suitable way the experimental set-up or observation requires) the hypothesis (phenomenon), then the hypothesis has passed the test.[12]

This is rather a roundabout way of bringing appropriate meaning to the associated requirement in standard analytical epistemology that I mention in Section 4.2 that for evidence to support a theory, the theory must entail the evidence. Another shortcoming of such logical analysis of the evidential relationship is that it fails to capture the intuition that prediction (of novel results) provides a more stringent test of a hypothesis than accommodation (of pre-existing data). This intuition seems to be common among the scientific community and many authors writing on the philosophy of science [13] have defended some form of such a "use novel" (UN) criterion for data in line with it. An initial approximation to such a requirement is that for a test of a hypothesis to be valid, it must quite literally be new; that is to say that novelty is a matter of temporal order. (This seems to have been the idea defended by Popper, for example.) It may sound strange to contemplate the idea of a test existing before the hypothesis that it is a test of; but in fact, as I have explained a test of a hypothesis here, this is a common occurrence in science: new theories can be developed precisely to resolve existing anomalies. The match between the novel theory and the pre-existing data that represented

[12]Note also that in this way we avoid talk of "proof" and can see evidence as accumulating and the necessity of opinion and judgement in possible cases of conflicting evidence.

[13]This view is clearly present in the writings of Popper, Lakatos, Giere and Worrall, to name just a few, and has given rise to a large volume of literature.

an anomaly can thus be seen as a test of the new theory. However, this notion of temporal precedence appears to be too strict a requirement, since what seems to be a stringent test may well be provided by a pre-existing problem, as in the case of GTR and the anomalous precession of the perihelion of Mercury, for example. Thus, a different interpretation of UN requires that for a test to be considered to fulfil a UN requirement and therefore not immediately be discredited, it must not form part of the "problem–situation" that led to the formulation of the hypothesis being tested.[14] On this reckoning, Mercury's anomalous perihelion advance would seem to count as a UN test, and therefore a valid test, of GTR; since, although the problem was well known, there is no suggestion that Einstein developed GTR in order to account for it. However, this requirement is clearly a very subjective criterion and requires knowledge of what a scientist had in mind when devising a theory!

A more sophisticated formulation is to disallow, as automatically lacking severity, the re-use of data that were used in the construction of a specific hypothesis as evidence in support of that same hypothesis. Thus we arrive at what seems to be a general and representative contemporary interpretation of a UN criterion (which in my experience is broadly seen as a requirement for any match between data and prediction to be counted as a real test, among physicists and philosophers alike), as espoused by John Worrall, for example, that both (to paraphrase Worrall):

> a hypothesis must entail the evidence;
> and the evidence must not have been used in the construction of the hypothesis.

The first part of this formulation is what I have already examined; and in contrast to its often impossible relationship of entailment, following the framework of Bogen and Woodward, I consider it more useful for my purposes here to see data as evidence for phenomena, rather than observations being in any way logically entailed by a hypothesis. So now I want to consider the second part of this UN requirement and thereby address two closely related worries concerning the modelling and interpretation of data: the possible legitimacy of non-novel use of data as a genuine test of what has been called

[14] This idea for a UN criterion is due to Zahar, 1973.

a use-constructed hypothesis, or double counting; and the (lack of) severity of tests of hypotheses.

Mayo (especially Mayo, 1991; Mayo, 2008, though also several others) has defended the idea that the UN intuition is only a partial expression of a broader epistemological requirement for severity in the tests that we subject hypotheses to experimentally. Indeed, the aim of introducing a UN criterion is certainly to disallow, as lacking severity, cases of double counting. The archetypal example of this practice is parameter adjustment in which the specific mathematical value of a parameter in a theory, which is unknown prior to observations, is set in accordance with the dataset observed.

If I may present another toy example: consider that I want to know how much petrol I need to put in my new car, now that the full tank it came with is empty, in order to be able drive 100 km. In this case, I know from the car's instrument panel that with the 100 litres of petrol the car came with in the (full) tank it had when I bought it, I have driven exactly 1,000 km. So from the data I have observed (1,000 km required 100 litre; or with 100 litres I drove 1,000 km) I arrive at a hypothesis that I will use or need (approximately) 10 litres to drive the next 100 km. When I am asked by a colleague how I can possibly know how much petrol I will need to drive the next 100 km, I very confidently say "because I worked it out from how far I drove with the last 100 litres". Should my colleague now criticise my reasoning and tell me that what I have arrived at is a totally speculative hypothesis and that I cannot have any faith in it as it has not been tested? I think not. This is an archetypal example of double counting in which the data serve both to formulate the hypothesis and as perfectly valid test of it: I can have faith in the veracity of my hypothesis because I have the data (that I used to construct it) to back it up.

In the face of such examples, which seem to be commonplace both in real life and in actual experimental scientific reasoning, Mayo maintains that although on occasion novelty of use may coincide with a severity requirement (maybe this is usually or even almost always the case), there are important ways in which non-novel uses of data can provide severe tests of hypotheses (while, of course, novel uses may fail to meet standards of severity in other ways). Thus

Mayo explicitly defends double counting in certain scenarios, in favour of ensuring the severity of the tests faced by hypotheses constructed and tested using the same dataset. So, as opposed to the UN criterion taken by so many to be a necessary condition for a severe test, Mayo introduces her requirement for severity in the following way, talking of whether the test, T, that the match between the data and the prediction made in accordance with the veracity of the hypothesis under consideration represents is passed:

> (SC): *Severity Criterion*: There is a very high probability that test T would not yield such a passing result, if T is false ...
> ...What SC is requiring is that there be a high probability that the test does not pass hypotheses erroneously. Probability is understood as relative frequency in a (real or hypothetical) series of test results.
>
> (Mayo, 1991, p. 529)

It is important to bear in mind that both views (SC and UN) developed as ways of evaluating how a particular dataset that is acquired experimentally can provide evidential support for a theoretical hypothesis (in terms of that hypothesis describing the way the world really is) via the hypothesis passing a stringent or severe test (ways that go beyond the standard epistemological approach). This characteristically Popperian idea that we need to be seriously trying to refute a hypothesis for any test that it passes to really count as evidence in its favour is what both the SC and UN criteria try to capture. To see how the two can diverge in practice, I will go on to consider a real example below. But before that I want to stress the importance that Mayo gives in her 2008 paper (Mayo, 2008) to the procedure by which a non-UN hypothesis is reached. That paper was written in response to criticism of SC and particularly in response to the criticism that double counting is an all or nothing affair which leads to a dilemma that Mayo sees as completely false (as set out in Hitchcock and Sober, 2004). The dilemma is that if we do not disallow double counting, then since every use-constructed hypothesis cannot fail to match the data, it is minimally severe (totally "insevere") in that whatever the data, we always arrive at a hypothesis that is supported by them. However, as Mayo goes to great lengths to stress, her SC is not limited to the idea that data must fit a hypothesis to a suitable degree; it is the likelihood of a use-construction procedure producing a hypothesis (that by definition

fits the data) even though the hypothesis is *false* that determines severity in these cases (and all others). The procedure or construction rule, R, by which we arrive at our hypothesis is explicitly assigned importance; and we need R to be reliable in order to come up with severe tests. This immediately ties in with the ideas of Woodward that I discuss in the previous section where he links counterfactual sensitivity to reliability. We need to follow the New Experimentalist path and reintroduce the action and judgement of the experienced and skilful experimenter who contrives the tests that experiments represent in such a way as to ensure that in these cases, not only do the data match the hypothesis (as they must in genuine cases of double counting) but that they would not have done so had the hypothesis not in fact turned out to hold.

Let me first just make it clear that this practice of double counting is certainly not one of the typical ways in which science advances. It is far more common to use initial observations (data) as evidence for an underlying regularity (phenomenon) and then to make predictions based on the supposed existence of that phenomenon. These are then tested either by making more observations (resampling the system and collecting more data) or by adjusting the system to see if the data collected afterwards track the adjustments in a way that is consistent with the effects the proposed phenomenon predicted; always ensuring the required counterfactual sensitivity through the specific set-up. That is, we use *novel* observations (data) to test our hypotheses. This is not always possible, though, as in cases of historical or one-off datasets, such as those that occur in much of geology, for example; or maybe it is not desirable, due to costs or risk factors. In cases where it is not possible or desirable, we often adopt a method that can be seen as trying to eliminate possible rival hypotheses through the gradual accumulation of indirect evidence, which increases our confidence in the use-constructed explanation. Thus, we may be left with a situation in which the data (observations) that led us to formulate our hypothesis are still the best evidence we have in its favour (and that evidence may indeed be compelling!), always tracked via the appropriate counterfactual sensitivity, as I have already indicated.

This is how much detective work advances: the evidence suggests a specific

scenario and through ruling out other possibilities, conclusions can often be reached that are backed only by the originally available (now non-novel) direct evidence. The crucial point is that for the hypothesis we arrive at in this way to pass a stringent test, we have to have constructed that test in a reliable way with respect to the hypothesis in question thereby ensuring that we have ruled out as many other possible explanations (rival hypotheses) as possible. This maintains the counterfactual sensitivity of the hypothesis so that it would not have passed the test if it were not correct.

It is also clear here, as Mayo indicates and we saw with Woodward too, that the stringency of a test comes in degrees (though we can consider—ideally—maximally and minimally stringent tests). Thus, in contrast to UN, a probabilistic SC requires of a test, T, that (to paraphrase Mayo):

> h is (highly) unlikely to pass T if h is false (or, h is (highly) unlikely to fail T if h is true)

where "passing" T means producing a result that fits h at least as well as e does.

Now, the example I wish to consider in detail of justified double counting, or non-novel use of data, is that of the detection of SN1987A as the first ever neutrino astronomy observation.

Neutrinos are electrically neutral subatomic particles of (approximately[15]) zero mass. They were first postulated as far back as 1930 by Pauli in order to balance the books, so to speak, and respect the conservation laws in the observations of beta decay. Fermi then completed the work of building them into our understanding of the beta decay mechanism. The 1995 Nobel Prize in Physics was awarded for the detection of the neutrino in 1956;[16] a sign of the long-lasting controversy typically associated with such detection. I will not go into the details of the solar neutrino flux, which constantly showers the

[15]The Standard Model of high energy physics requires the mass of the neutrino to be zero; however, this is still an extremely hot topic, as any switching of neutrino "colour", which apparently has been detected, would theoretically require the neutrino to have some mass. Meanwhile, certain theoretical considerations together with observations, including that of SN1987A, place a very small upper limit on that mass.

[16]See Cowan et al., 1956, for the original paper reporting the discovery.

Earth in neutrinos of solar origin; or the difficulties associated with detecting neutrinos, which usually pass straight through terrestrial matter. Suffice it to say, that it is extremely difficult for us to detect (and hence count) neutrinos: it requires large (hundreds or thousands of cubic meters), carefully shielded (usually placed in mines, well below the Earth's surface) detection tanks surrounded by (usually multiple layers of) scintillation detectors and photomultiplier tubes. Even so there have been several such functioning detectors.

On 24th February 1987, a new supernova, SN1987A (as it was called since it was the first supernova detected that year), was observed in the southern sky and located within the Large Magellanic Cloud. At the time, 4 neutrino detectors were operative around the world. To the best of our understanding, some supernovas (Type I) do not produce neutrinos; while the others (Type II) do. The team at the Kamiokande neutrino detector in Japan is reported to have received the breathtaking news of the supernova sighting thus: "Supernova went off in Large Magellanic Clouds. Can you see it? This is what we have been waiting 350 years for!"[17]. (The reference to 350 years alludes to this being the closest supernova to the Earth since the 1604 supernova in our own galaxy known as Kepler's Supernova or Kepler's Star, since he is known to have dedicated time to observing it.) It was unknown whether SN1987A was a Type I or Type II supernova, but within days, analysis of the data from the Kamiokande detector[18] confirmed a "cluster" of 11 neutrino events (compared to a "negligible" theoretical background detection rate of 1.2×10^{-8} per year coming from the rare solar neutrinos with sufficient energy to be detected) over 45 s, at 0735 h UT on February 23rd.

Two hypotheses were thus formed based on these observations and they

[17] See Totsuka, 1991, p. 524

[18] The 3 other detectors also registered clusters, but one yielded a hitherto unexplained timing anomaly of several hours while the other two detected fewer events, so I concentrate on the Kamiokande data. That there were 4 detectors in total does not change the analysis of the combined dataset as both the origin of the hypotheses and their confirmation, in exactly the same way as GP-B carried 4 gyros. However, in both cases, if the actual data collected by the four separate set-ups were consistent with similar hypotheses derived from that data in each different cases, then we must see this as increasing our confidence in them.

also underwent what I see as stringent tests: they meet Mayo's probabilistic SC. First, that SN1987A was a Type II supernova; and second, that this was the first observation in the history of neutrino astronomy. Assuming that the Standard Model of high-energy physics is a reasonable description of how things work at the subatomic level, we can track the truth of these findings via the appropriate counterfactual reasoning. In the first case, had SN1987A been a Type I supernova, no cluster of neutrino events would have been detected. In the second, making the probabilistic nature explicit: it is extremely improbable (to the point of being virtually impossible!) that had SN1987A not produced the neutrinos which resulted in the events detected by Kamiokande (that is, those events were in fact some type of multiple malfunction; or resulted from some other extraterrestrial event or even some hitherto unknown and undetected terrestrial cause) that they would have coincided with the observation of SN1987A to such a degree.

I hope the case is clear here that a one-off, unrepeatable event led to the collection of a specific dataset. That dataset was then used to arrive at certain hypotheses. Those hypotheses were confirmed by their agreement with the original dataset. This is only possible due to the tracking of actual difference makers via the counterfactual reasoning I have briefly laid out. The conclusions are nonetheless considered valid scientific findings; and they are considered to have undergone stringent tests (matching the dataset) despite this double counting.[19]

[19] A further real-life example was related to me by a colleague who, suffering from alopecia, was told by his doctor that he was suffering from stress. On questioning the diagnosis, the patient was assured that it was correct as his alopecia was a clear sign of stress. Alternative causes had been suggested and satisfactorily ruled out via the appropriate counterfactual reasoning—had his condition been due to a viral infection, he would be suffering from some specific additional symptoms; if it were a consequence of some known kind of intoxication, this would have shown up in the analysis of a recent blood test; etc. So the conclusion, the evidence for which was precisely the observation that had led to its formulation, was that the patient was suffering from stress, apparently unaware of it and with no other symptoms. (In this case it is easy to see that the rigour of the argumentation may not be what we would desire, but nonetheless, it shows that it was apparently sufficient for certain ends.) Other examples I have considered include the Tunguska Event on 30/06/1908 and the Oklo natural nuclear reaction phenomenon. Certainly in the latter case, and maybe in the former, a unique dataset led to the formation of a hypothesis which has

To reiterate: the probabilistic SC states that the chance of the evidence (data), e, fitting (in whatever way is required) our hypothesis, h sufficiently well for us to consider the test passed and therefore h confirmed, when h is false, must be (very) low. The inclusion of "when h is false" is vital and what seems to be missing from the UN criterion; it is not necessary (and it certainly is not sufficient) for e not to have been used in constructing h for such a severity criterion to be fulfilled. Mayo defines severity thus:

> A test's severity is one minus the probability it yields some such passing result (or other), given the hypothesis passed is false. (Mayo, 1991, pp. 530-1)

A final comment is in order regarding the relative positions of Woodward (and Bogen, to a much lesser extent) and Mayo. While Woodward's position in the papers he has published is very cautious and he stresses that his "goal was to advance a more pluralistic understanding of science" (Woodward, 2011, p. 171), Mayo's goal in her books seems to be the far more wide-ranging programme of setting in motion a "non-Bayseian philosophy of science that may be called the error-statistical approach" (Mayo, 1996, p. 442). So while they seem to coincide in the details that I have put forward here, it seems clear that the scope and overall aims of their work are different. Woodward has the following tantalising introduction to further work by Mayo on the subject:

> . . . it is one thing to talk of the error characteristics of Eddington's procedures for inferring from his photographs to a value for the starlight deflection and another matter to talk about the error characteristics of some procedure involving the use of starlight deflection to test General Relativity. Does it makes sense to think in terms of a repeatable evidence generation or testing procedure with determinate empirically accessible error characteristics in the latter case as well as the former? As I understand the argument of Deborah Mayo's recent book (1996), her answer is "yes"; she wants to appeal to ideas about error characteristics to give an account of testing and evidence in science which is applicable quite generally and not just in the context of data-to-phenomena reasoning. While I share Mayo's emphasis on the importance of error characteristics in the context of data-to-phenomena reasoning, I confess to some

been considered confirmed, based precisely on that dataset: all rival theories have been successfully ruled out thanks to indirect evidence.

> uncertainty about how she proposes to extend these ideas to other contexts.
>
> (Woodward, 2000, p. S172)

I hope to have shown here my reasons why I see the analysis of the counterfactual relationships between experimental evidence (data) and target hypotheses (describing phenomena of interest) as propounded by Bogen and Woodward, combined with a probabilistic severity criterion of the type advocated by Mayo, as the best method to capture and describe the reality of some if not most scientific experimentation. It is not just a useful description that is meaningful to scientists and reflects their concerns, it is also a prescription concerning how to establish the required scientific link between (observed) data and (often unobservable, and certainly more theoretical) phenomena through the identification of genuine difference makers. Furthermore, such an approach sets out clear aims and goals for test passing that reflect epistemological requirements but avoid over-simplistic or rigid criteria that scientists (who cannot be expected to remain abreast of theoretical advances in philosophy) often seem to adopt.

In the next chapter (Chapter 5) I give some details of the actual GP-B results. There, I review the novel methods adopted by the team to save the experiment, in the face of the excessive noise I have mentioned previously. That is before moving on in 6 to use the analysis I have developed here in this chapter, combining the work of different the authors I have introduced in some detail here, as the basis of the method I apply in a large part of my analysis of GP-B.

Chapter 5

Gravity Probe B Science Results

"Direct tests of nature's laws are the foundation of physical science; such tests are the only rational basis for the belief that these laws are, in part, 'understood.' GP-B seeks to deepen our understanding of gravity in this way."

Bob Kahn, *GP-B Mission Update: 26/09/2008*

5.1 Introduction: Trouble at t'Mill[1]

As soon as the first GP-B data were collected and analysed during the Science Phase of the space mission, it became clear that the gyros were not behaving as expected. Despite this, the mission went ahead almost as planned and concluded successfully in terms of the volume of data collected. However, the post-flight analysis of the data was dogged by problems. As I have already said in previous chapters, those problems led to the data analysis period being extended and new innovative methods being developed to attempt to recover valid results from the unexpected and unwanted effects that threatened to swamp completely the relativity signals that were supposedly buried in the extremely noisy dataset the team had collected. Prior to the final preparation for the launch of the mission it had always been hoped that the results would refine the limits on the PPN parameters that determine the agreement between

[1]For the reader who is not familiar with this comic British expression, this is part of what Oxford Dictionaries online has to say about it: "a humorous phrase sometimes used by British people to refer to a problem, especially at home or at work."(*Oxford English Dictionary*)

GTR and observation within a weak-field low-velocity setting. However, coupled with findings from other sources that had been published since 2003, the problems in arriving at new limits on the PPN parameters from the noisy GP-B dataset appeared so insurmountable that in 2008 they led to NASA ending its funding of the ongoing GP-B data analysis. That left the Stanford team effectively on its own to attempt to raise the funds it deemed necessary to complete its new data analysis strategy and arrive at final results.

The decision by NASA to withdraw funding, or at least not to renew it, was taken in May 2008 after Francis Everitt was (surprisingly, given his history of resurrecting the project from the ashes on previous occasions[2]) unsuccessful in his personal defence of the proposal submitted in March of that year to the NASA Science Mission Directorate, Astrophysics Division Senior Review of Operating Missions. In that proposal, the GP-B team requested a final 18-month funding extension to run through March 2010. During the 7 or 8 months immediately prior to the end of NASA funding, after 15th January 2008, financing for the continuation of the project had come from a private donation which had been negotiated as a stopgap and matched by similar amounts from both NASA and Stanford University.[3] Previous information regarding funding published in June 2007 stated that, in accordance with the

[2]Recall that in Chapter 3 footnotes 15 and 16, I recount how the entire GP-B project had nearly been scrapped 7 times previously over the course of its history and every time personal intervention by Everitt had resulted in those plans being revised and the project continuing, with the space mission eventually going ahead in 2004-5.

[3]In March 2008, the mission update informed us:

> "On November 2, 2007, we convened the 17th meeting of our external Science Advisory Committee (SAC) to review our progress in the refinement of the GP-B experimental results. The subsequent SAC report noted 'the truly extraordinary progress that had been made in data analysis since SAC-16 [March 23-24, 2007]' and unanimously concluded 'that GP-B is on an accelerating path toward reaching good science results.'
>
> Following a peer-reviewed bridging proposal to the NASA Science Mission Directorate (SMD) and actions by Stanford and a private donor, the GP-B program has been extended at least through September 2008. Furthermore, SMD opened the opportunity for GP-B to submit a proposal this month to its Senior Review process. This is a bi-annual event in which ongoing NASA science programs undergo a peer-review to determine which of those programs NASA should continue and/or extend in order to achieve the greatest scientific gain. Assuming a successful Senior

March 2007 recommendation by the GP-B Science Advisory Committee that funding should continue through December 2007, NASA had "committed to extending support for GP-B at the required level." (Kahn, 2007, p. 427)[4] In the event, the temporary bridging proposal came into effect on 15th January 2008. Everyone on the GP-B team seems to have been surprised that after almost 45 years of financial backing, NASA did not agree to continue funding the project beyond 30th September 2008.

Eventually, in the Status Update of 26/9/08 published as always on the project website, Bob Kahn informed readers that

> GP-B has secured alternative funding that will enable our science team to continue working at least through December 2009 in order to complete the data analysis and bring GP-B to a proper close.
>
> (*GP-B Mission Update: 26/09/2008*)

That funding came from the King Abdulaziz City for Science and Technology (KACST) in Saudi Arabia, with which Stanford University thereby set up an important collaboration. The university appointed Professor Charbel Farhat of the Stanford University Aero-Astro Department as Co-PI for GP-B data analysis. That left the situation NASA was in with regards to any later results extremely unclear. The Administration had funded the project for over 44 years, but apparently would not be part of the team publishing final results, supposedly just 2 years after they pulled the plug on funding. In the event, it was nearer 3 years later that final results were announced: in May 2011, at a press conference organised and held jointly with NASA.

Once again, this episode illustrates just how important major government policy decisions (via NASA, in this case) and indeed international politics

---

Review, GP-B will be extended one final time, from October 2008 through March 2010."

(*GP-B Mission Update: 21/03/2008*)

[4] This comprehensive, 600-page 2007 report was clearly written as a major GP-B team collaborative effort. However, as with the later Kahn, 2008 I mention in footnote 7 below, the document indicates that it was "Prepared by Robert Kahn" as the Public Affairs Coordinator and the team leaders endorsed it. I have therefore similarly referenced it throughout this work as Kahn, 2007.

(the collaborative effort between an emblematic US university, albeit a private institution, and an initiative patronised by a notoriously brutal and repressive totalitarian regime) can be to fundamental physics experiments.[5] Blue-skies research, such as GP-B, with no clear practical goal or marketable end product has become harder to find funding for as enthusiasm for what is often called "Big Science" has dwindled since the mid twentieth century during the post-WWII economic boom period and before the energy crisis of the 1970s. These days, although Big Science projects are certainly alive and well around the world, and indeed more and more as international collaborations rather than purely instruments of competition between nations and blocs, the growing importance of private investment in science research has meant that the potential for profitable spin-offs and a monetary return have taken on a higher profile and moved up the list of desirables for projects looking for investors. As a consequence, basic research has become less attractive and more difficult to find funding for. Within this contemporary setting and as a public entity, NASA, often acting as a partner with private interests such as Lockheed Martin, as in the case of GP-B, has to respond and react to mixed and often contradictory pressures and influences. Once the Science Phase of the space mission was over and the extent of the problems with the GP-B data emerged, a prolonged period of number crunching and theoretical work became necessary to bring the project to closure. The only available document that explains NASA's decision is the Astrophysics Division Senior Review Committee report from 2008,[6] which dedicates just 2 pages to GP-B and leaves exactly how different considerations affected the NASA decision to end funding of the project largely to speculation.

It is interesting here to contrast the criticism coming from that NASA Senior Review Committee report, which here I will refer to simply as the NSRC

[5] Again, recall that GP-B originally got off the ground, so to speak, in the midst of the Cold War and it seems to have been supported by military interests within the USA.

[6] This 2008 NASA report is an unsigned document. It indicates the members of the Committee and names the chair as Richard McCray; however, as no specific author is indicated, I have referenced it simply as NASA, 2008

report, with the *Science Results—NASA Final Report*[7] dated December of that same year, 2008, which was a report to NASA by the GP-B team and which here I will refer to just as the Science Results report, for clarity. The Preface to the Science Results report informs us that, at the time (2008), the experiment

> yields a consistent determination of the geodetic effect to 0.5%. The frame-dragging effect is plainly visible in the processed data with a present statistical uncertainty of 15%.
>
> (Kahn, 2008, p. 1)

This is then heralded as an important result, but the lack of importance of this result was precisely one of the factors cited in the NSRC report to explain why NASA had decided not to continue funding the data analysis. The NSRC report claimed that other measurements had effectively leap-frogged such a result from GP-B and surpassed these determinations in accuracy. We are told that:

> the GP-B experiment has been somewhat overtaken by events and now occupies a diminished niche in the field of experimental tests of GR. Theories that can be expressed in Post-Newtonian (PN) formalism in the weak-field limit are sharply constrained ... [current limits on the PN parameters] are about 2 orders of magnitude below the limits that might be achieved by GP-B in the team's optimistic projection of what they can do.
>
> (NASA, 2008, p. 24)

The doubts expressed here by the NSRC were twofold: that the levels of accuracy that the team could hope to achieve were insufficient to be of interest; and that the team was overly optimistic in thinking that it could refine its results even to the level that had originally been the target (already surpassed by other measurements). I should note in passing that these claims by the GP-B team did indeed turn out to be overly optimistic as the final result arrived at and published 3 years later gave an accuracy of $\pm 18\%$ on the frame dragging: below the degree of accuracy that the team was claiming back in 2008. The accuracy that the NSRC report refers to as being two orders of

[7] This 2008 report was clearly written by the GP-B team as a collaborative effort. However, the document indicates that it was "Prepared by Robert Kahn" as the Public Affairs Coordinator and the team leaders have endorsed it. I have referenced it throughout this work simply as Kahn, 2008

magnitude better than the tightest restriction that the GP-B team were even aiming at is that provided by measurements using the Cassini spacecraft. Cassini reached Saturn and went into orbit around it in 2004. Using signals sent during the Cassini mission, the limit on $|\gamma - 1|$ was calculated as $< 10^{-3}$ and this was later refined to as little as $< 10^{-5}$.

However, due to the fact that it was an experiment, and to the specifics of the experiment, the team were able to claim that it was still the most accurate experimental determination of this parameter to date that did not rely on massless particles (photons) and therefore an extremely valuable milestone in experimental physics. The team also emphasised other qualitative differences between GP-B and other measurements. As I mention in Chapter 2, footnote 25, Overduin claims that:

> The importance of the geodetic and frame-dragging effects, however, does not lie in their implications for the PPN parameters. It lies in the fact that both these phenomena are qualitatively different from all preceding tests of GR, in that they depend on the spin of the test body and/or source of the field.
>
> (Overduin, 2015, p. 4)

However, to go back to 22nd – 25th April 2008 when the NSRC met, irrespective of the other results the committee mentions and possible arguments concerning the relative merits of each, it seems clear that there were real and well-founded doubts as to the validity of the processes to be used by the GP-B team to arrive at their final results. The review board made it clear, as can be seen in the quote above, that they (correctly, as we now know with hindsight) considered the GP-B team's hopes and appraisal of its own capacity to continue to refine their provisional results and produce reliable more accurate results to be at the very least "optimistic". This is the only published criticism of the GP-B project and the claims concerning the results that I have found. However, my personal interaction and communication with physicists at the time made it clear to me that there was certainly a trend of opinion, if not a widespread conviction, that the project had failed to deliver on its goals and could not be saved.

As I have already made clear, in the light of the problems and before NASA withdrew its funding, the members of the GP-B team had set about devising a

new analytical method that they hoped would allow them to separate out the observations of the two GTR effects from the unwanted disturbances that the GP-B gyros suffered. In a comment on this situation in the Preface of the 2008 Science Results report, we are told that the data analysis was "more subtle than expected." Considering the ongoing extension of the data analysis phase of the mission from the expected 10 months or so (at launch up to one year was allowed for the post-re-calibration period of data analysis) to over 5 and a half years (October 2005 to May 2011) and the remarkable complexity of the methods devised in order to attempt to salvage results and credibility for them, this would seem to be a considerable understatement. In fact the NSRC report specifically states that one of the weaknesses of the revamped GP-B proposal was that:

> any effect ultimately detected by this experiment will have to overcome considerable (and in our opinion, well-justified) skepticism in the scientific community.
>
> (NASA, 2008, p. 25)

As the only published material on this matter, it is this literal expression of the worries and scepticism concerning the results that I will take as the starting point for my further examination of the GP-B data analysis process and the team's eventual claims.

In my efforts to discover more of the motivation for these brief comments, which as I have said were indicative of the reasons for bringing to a close over 44 years of NASA funding of GP-B, I later spoke to Richard McCray, the Chair of the NSRC. In our private communication, McCray confirmed to me that one of the main concerns of the NSRC was that with large systematic errors affecting the final dataset, whatever the process by which those errors were removed, there would have to be a significant degree of doubt concerning the validity of that removal process. Part of the problem here, as I see it, is the worry that with a unique, unrepeatable dataset which needs to be cleaned up, there is no independent justification of that process. As the Chair of the NSRC, McCray also indicated to me that another member of the Committee, Neil Cornish, had probably been the one with the most relevant expertise concerning GP-B. Also in private communication, Cornish expressed a certain degree of surprise that no analysis criticising GP-B had been published. He

also explained that the criticism and scepticism was based more on general scientific principals and accrued expertise than on any specific reports (which they did not have at their disposal, as there do not appear to be any). Along these lines, the crux of Cornish's criticism, while praising the pioneering work of the GP-B team, was that the removal of systematic errors that are far larger than the target signal[8] is something that scientists just cannot be confident about. He suggested that one horn of the dilemma such a situation leads to is that if the final results agree with theory (the predictions of GTR in this case) then the clean-up process may well be open to criticism regarding the worry that the process would always have continued until such agreement had been reached.

These worries and criticisms form the basis of what I go on to consider and analyse in the rest of this work. In the next section, I consider exactly what the two anomalies in the science data consisted of. Then, in Section 5.3, I explain the underlying cause that the GP-B team discovered and consider their confidence in this finding before going on in Section 5.4 to discuss the solution proposed by the team. In the final section of this chapter, Section 5.5, I then analyse the extent to which this can be seen as valid science.

5.2 Anomalous SQUID Readout

5.2.1 Jumps in spin axis orientation

As I have already hinted at, GP-B experienced certain problems during the IOC phase of the space mission. This was the period when all the on-board systems were brought on-line, the satellite was manoeuvred into the correct position and orientation, crucially the gyros were spun up and aligned with

[8]The errors that were far larger than the effect that was the target of the mission were the relatively sudden jumps that occasionally occurred in the spin axis orientation of all the gyros; which I describe below. In the event, these sections of data were discarded and not included in the analysis. While this is clearly not an ideal solution, it does seem to counter this specific criticism; while possibly opening up the process to different criticism: that of discarding undesirable data and only using those data that agree with expectations!

the guide star, and the systems were calibrated and tested. In their Post Flight Analysis report of 2007, the team reported that it was found to be impossible to fly the satellite in its primary drag-free configuration. In this "free-floating" mode, one of the gyros was used as the "proof mass" and the satellite was piloted, so to speak, to follow the orbit of that gyro while gently adjusting (using the electrical suspension system) the other three gyros to follow the path of the proof mass. The gyro used as the proof mass was thus literally floating with no effort at all being made by its gyro suspension system. Instead, using an incredibly intricate and complicated feed-forward system, the satellite was flown around the proof mass and minute forces were exerted by the other gyro suspension systems on the remaining three gyros to maintain them in the centre of their housings. (The most important effect that needed to be balanced by the gyro suspension system on the other three gyros was the Earth's gravity gradient along the length of the SIA! This gives some indication of the tiny scale of the effects.) It was eventually decided to perform the entire Science Phase of the mission in secondary drag-free configuration. This is also called accelerometer drag-free flight as the proof mass too has some force exerted on it to maintain it at the centre of the housing, in such a way as to minimise the forces exerted on all four gyros together. References to this decision seem to be very rare in the papers comprising the 2015 Focus Issue of Classical and Quantum Gravity. There, the talk is almost always of the "nominal" drag-free operation. Although extensively tested on orbit, the primary mode proved to be problematic. It is necessary to go back to 2007 to find the following declaration concerning one of the failed attempts to switch over from secondary to primary drag-free mode during the Science Phase:

> Preliminary analysis indicated that this failed switch-over was similar to ones that occurred during IOC and was probably due to "un-modeled" forces between the gyro rotor and its housing.
>
> (Kahn, 2007, p. 64)

This is also another example of the caution taken to ensure a working mission; in this case an alternative rather than pure redundancy produced by duplication.

So it was certainly clear that here were problems right from the start of the space mission. However, secondary drag-free flight mode was used for

the entire Science Phase and it was believed that the loss of purity in the gravitational nature of the paths of the gyros as a result would be negligible. Here is the only account of this that appears in the 22 papers of the Classical and Quantum Gravity Focus Issue:

> Though free-floating drag-free is the preferred operational mode, it was not used during the science data-gathering phase of the mission. After on-orbit calibration, it was found that some gyroscopes exhibited a small acceleration bias, up to 20 nN, that the free-floating drag-free system would track. Though this bias would have no immediate effect on the drift performance [of] the gyroscopes, the ATC control action over time would slowly change the space vehicle's orbit. The accelerometer mode system could properly compensate for these biases and showed acceptable performance for the mission during testing, thus it was selected as the baseline drag-free mode during science data collection.
>
> (Bencze et al., 2015, p. 21)

During the Science Phase though, as I have said, the signals received from the SQUID readout were simply not as expected.

The readout was noisy to a degree that just had not been anticipated. The spacecraft appeared to behave as expected and there were no signs of instrument or system malfunction, but the SQUID readout from the gyros was excessively noisy. In many ways this was not the greatest worry though. At certain (initially irregular and unpredictable) intervals, the rotor orientation jumped from its original direction to a new direction by up to some 100 mas. This was no subtle drift but a sudden jump that would take a few days to complete, but would result in massive (by the scales of the experiment) shifts in alignment.

This can best be seen in an illustration. Figure 5.1 shows the actual data, in the form of the north–south component of the gyro spin axis orientation, collected for one of the gyroscopes over some 75 days between April and July 2005. As can be seen, there are 8 clear (and a ninth indicated) jumps in orientation, by as much as approximately 100 mas on three occasions, lasting of the order of 4 or 5 days each. In fact, as the vertical (numbered) lines in the figure indicate, as the mission progressed the jumps turned out not to be so irregular and that is part of the key to understanding their cause. The data

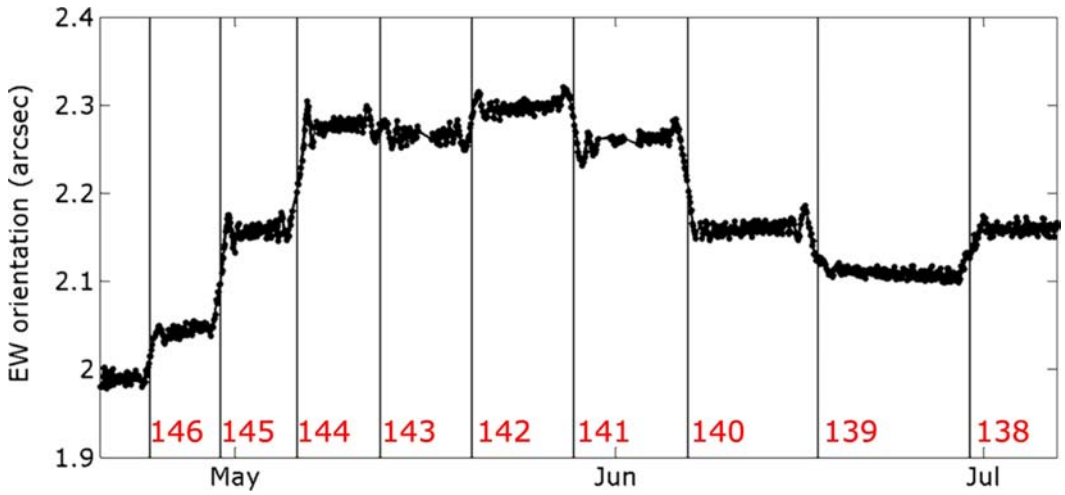

FIGURE 5.1: East-west component of the SQUID readout data over 75 days between April and July 2005, for a GP-B gyro. A total of 8 jumps can be clearly seen in the readout. Together with a ninth, they are indicated by the numbered vertical lines. [SOURCE: Everitt et al., 2015]

from a different gyro show the combined north–south and east–west shifts in orientation: Figure 5.2. (Recall that the data points that indicate overall gyro spin axis orientation are the sum total of all the information collected over one approximately 60-minute-long GSV period, every 98 minutes: some 15 data points per day). As can be seen here, the gyro seems to spiral out from its initial orientation, then skip the 100 mas or so and then spiral back into its new orientation.

Despite the problems encountered in the actual execution of the GP-B flight mission, it is important to note that as of the Science Results report to NASA in December 2008, the team still held to the original aim of determining relativistic drift of the four gyroscopes to <0.5 mas/yr. What had changed was the method to be used in order to achieve this degree of accuracy, which, if achieved, according to Everitt et al. would still mean that:

> GP-B would become the most rigorously validated of all tests of Einstein's theory. That aim, though attained by different means in the actual as against the ideal GP-B, is one we still hold to.
>
> (Kahn, 2008, p. 11)

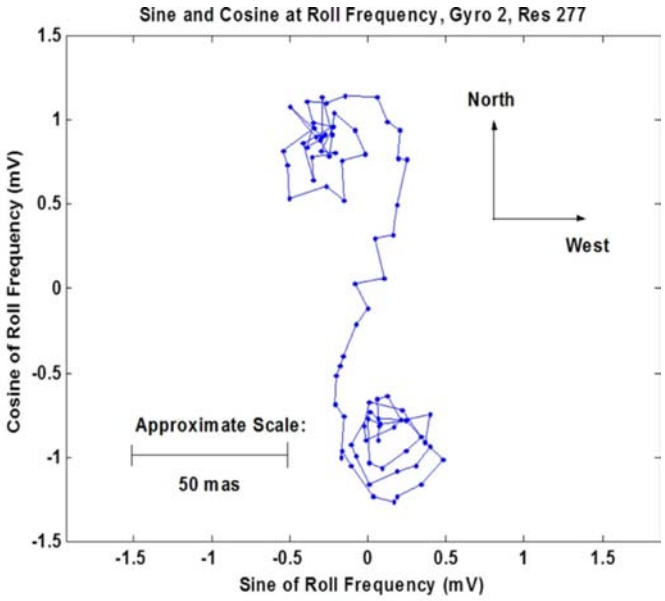

FIGURE 5.2: Both the east–west and the north–south displacement of one jump in the the SQUID data of a GP-B gyro, over about 5 days. The overall event can be seen to alter the spin axis orientation by as much as 100 mas in the north–south direction.
[SOURCE: Everitt et al., 2015]

This optimistic stance was partly due to the massive redundancy built in to the entire mission, and also to the fact that most of the on-board systems performed at least as well as required for the relativity signals to be detected. Those systems vitally included dewar functioning, telescope pointing and SQUID readout noise levels. The GP-B team leaves us in no doubt that the spacecraft and the readout systems functioned at least as well as expected, thus allowing the gyro orientation to be calculated from the combination of the SQUID readout and the spacecraft pointing information; despite those readings being so far removed from their expectations. It also seems totally reasonable for anyone, seeing such glaring unexpected errors in the data, occurring often and over a few days, and being larger than one of the effects that the experiment was aiming to measure, to be more than a little sceptical of

the possibilities of removing such errors from a unique, unrepeatable dataset![9]

5.2.2 Source of noise

These seemingly damning anomalies threatened to thwart the entire mission (and for the sceptics, indeed they did). In their characteristic, wonderfully understated manner when referring to this situation, Everitt at al. tell us that "some deviations from the ideal did occur" (Kahn, 2008, p. 11). As we have seen (optimism in hand, or to heart) far from throwing in the towel, the team set about trying to establish what was causing the anomalies: the excessive noise on the one hand and the jumps in gyro orientation on the other. The explanations that the team eventually came to favour were first suggested during the final calibration phase. Here the spacecraft was deliberately pointed away from the line of sight to the guide star to examine the effect on the SQUID readout. Furthermore, accelerations were introduced to the spacecraft, and the voltage supplying the gyro suspension units was adjusted both in magnitude and modulation, which was changed from a 20 Hz alternating current to 200 Hz and reduced to zero; that is, to a direct current supply. The large variations in spacecraft spin axis orientation produced much larger than expected drift rates, or torques, on the gyros and the changes in the gyro suspension system electrical supply provided crucial evidence that the origin of the torques was of an electrical nature. The switch from an a/c to a d/c supply resulted in a change in the scale of the misalignment torque though not in a change in the direction of the torque. As the steady voltage was increased or decreased so the degree of misalignment torque changed. Furthermore, the linearity in the relationship between the torque and the misalignment which held at small angles broke down as the misalignment increased. This can be seen in Figure 5.3 which shows precisely the strength and orientation of the torque on the gyroscopes as the satellite was pointed away from the guide star, that is: as the misalignment between the telescope pointing angle (indicating the guide star) and the gyro spin axis orientation increased. These findings

[9] Recall again that nearly 2 decades earlier Cartwright had precisely heralded the set-up as an unprecedented attempt to shield an experiment from effects other than the target, GTR, effects. On that count, it definitely seemed to have failed.

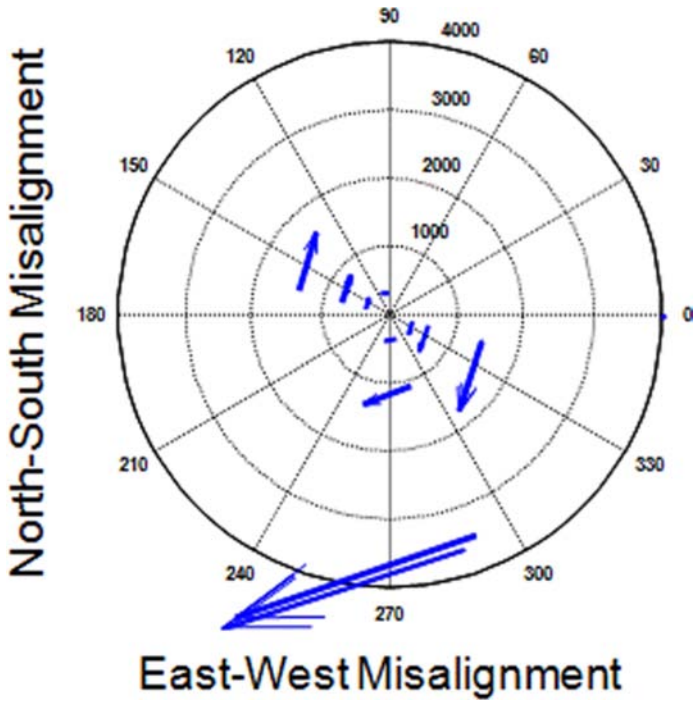

FIGURE 5.3: Misalignment torques on the gyros during calibration were seen to be larger than expected. It was also observed that they were proportional to the misalignment angle only for very small angles, not when a larger misalignment was deliberately introduced. Together with the dependence on the frequency of the a/c power supply this was evidence of the electromagnetic nature of the anomalies.
[SOURCE: Everitt et al., 2015]

were crucial as they provided the team with a point of entry or the starting point they needed if they were going to find ways of tackling the apparently overwhelming problems they faced.

It was the magnetic London moment (LM) of the gyros—generated by their spinning superconducting shells—that the on-board SQUIDs were supposedly measuring. The LM would induce a current in the SQUID pick-up loop as the latter slowly rotated around the gyro at the spacecraft roll rate. However, any additional electromagnetic field associated with the gyro could form an additional component that would also induce a current in the pick-up loop and hence be detected by the SQUID. As I have just explained, and

as was exploited by the team during the post-science calibration phase, any misalignment between the gyro spin axis orientation and the spacecraft spin axis orientation would result in a net torque on the gyro. Furthermore, that torque would tend to make the gyro precess about the spacecraft spin axis orientation. To reduce such gyro precession, the GP-B experiment was designed so that over the whole mission the gyros pointed as closely as possible along the spacecraft spin axis orientation; that is, the misalignment between gyro and spacecraft spin axis orientation is minimised in the experimental design. Precisely what the experiment aimed to measure was a shift in the gyro spin axis orientation (towards south and east, in standard Earth coordinates). But any additional electromagnetism of the superconducting shell that was not perfectly aligned with the spacecraft roll axis would also cause larger than expected Newtonian torques on the gyro.

Since it was expected that the superconducting gyros would have a perfectly even charge distribution, there would therefore be no overall systematic electrostatic effect on the spin axis orientation, and the only net torque due to misalignment would come from the LM. For the tiny misalignment involved (a range over which the relationship between gyro spin axis orientation and spacecraft spin axis orientation can be approximated as linear) the predicted torques would be negligible compared to the relativity signal (plus the effects of the Earth's and the Sun's magnetic moments, and the Earth's oblateness; together with the Sun's geodetic effects). Any residual torque that affected the gyro motion was expected to be periodic at the (constant) spacecraft roll rate or a constant frequency effect known as the gyro polhode motion, which I explain below. Such effects would all average out over their respective frequencies (the spacecraft was rolling once every 77.5 seconds; the gyros spinning at between 5,000 and 9,000 rpm) thereby producing no effect on the science result and they would be accounted for in the calculation of the gyro scale factor which was empirically derived during the mission. However, the evidence from the post-science calibration phase indicated that there were larger than expected electromagnetic forces causing torques on the gyros and these did in effect cause larger than expected Newtonian torques on the gyros (precisely one of the possible interferences that was to have been shielded against). This is what was seen in the SQUID science data readout

as excessive noise. Furthermore, the data analysis that had been planned for crucially included calculating the scale factor that would allow the voltage reading that was the science data, the output provided by the SQUID that detected the LM, to be converted into an angle. As I have just mentioned, the team had expected the polhode motion of each gyro to be constant over the lifetime of the experiment, but possible interactions with local electromagnetic fields made them question that assumption. This effect known as polhoding or nutation was to prove to be the key to unravelling the anomalous GP-B data and so it is worth taking a moment here to explain it.

5.2.3 Polhoding

Polhoding is the motion of a (solid) spinning body around its spin axis. It is important to note that polhoding is not a perturbation of the spin axis orientation relative to inertial spacetime (which would be a precession), but rather it is a perturbation of the motion of the spinning body itself around its spin axis. It is the result of deviations from perfect spherical (or cylindrical) symmetry: perfectly spherical bodies experience no polhoding. Something very similar can be visualised in the "rising egg" phenomenon, whereby if we lie a hard-boiled egg (or rugby ball) horizontally on a surface and spin it fast enough, it tends to rise up and "stand" on end (or vertically). Figure 5.4 illustrates how a hard-boiled egg tends to move in this fashion. It is critical to appreciate that while the geometric axis of the egg (what we can consider as the longest straight line segment passing through the egg) moves considerably (from almost horizontal to almost vertical), its spin axis continues to point in exactly the same direction. Any further lack of (cylindrical) symmetry will then cause the egg to bob up and down by tiny amounts, or wobble, as it shifts its position relative to its spin axis thereby minimising its moment of inertia as it continues to spin.

To see how this affects the spin motion of an almost perfect sphere (such as a GP-B gyro), we need to consider its principal axes. For any solid body, we can define three principal (passing through the centre of mass of the object) axes: the major principal axis, about which the moment of inertia is

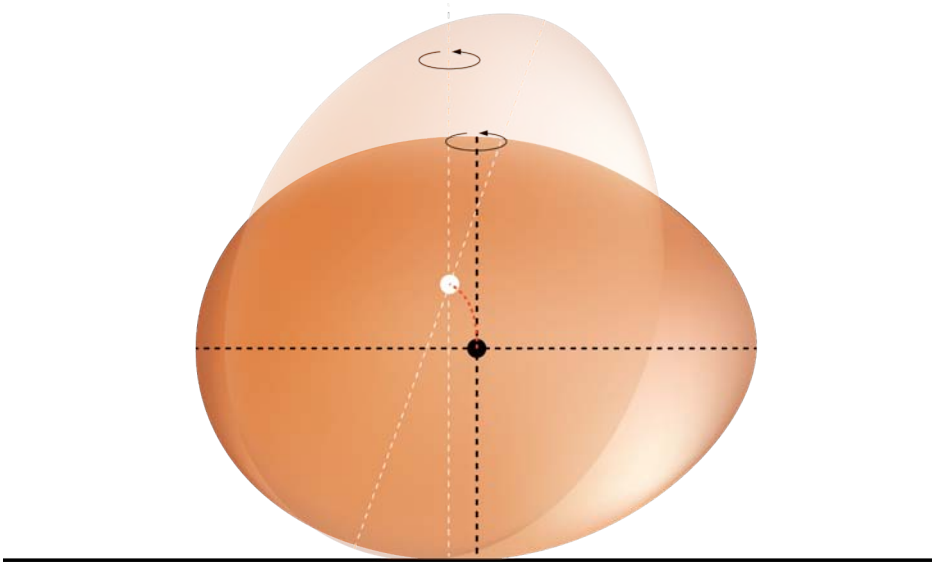

FIGURE 5.4: If a horizontal hard-boiled egg spins fast enough, it "rises up" to the vertical, thereby reducing its moment of inertia about its spin axis. In the case illustrated here, due to friction with the surface, there is a slight shift in the centre of mass and in the spin axis through it; but—crucially for the possibility of solving the problems with the GP-B data—there is no change in the orientation of the spin axis. We can envisage that if the egg were floating, it would simple rotate about its centre of mass which would remain fixed. Even though the axis of geometric orientation of the egg (the longest axis through the centre of the egg) is seen to change direction greatly (from horizontal towards vertical), the spin axis remains pointing strictly up. Deviations from perfect (cylindrical) symmetry will then cause the axis of geometric orientation of the egg to bob up and down as it spins, thereby maintaining its moment of inertia constant. Throughout this periodic motion, the orientation of its spin axis remains fixed. [SOURCE: *Perimeter Institute for Theoretical Physics*, ©2012]

greatest; the minor principal axis, with minimum moment of inertia; and the intermediate principal axis perpendicular to the previous two. If, as in the case of the GP-B gyroscopes, we spin a nearly spherical body around an arbitrary axis, then the body will tend to wobble slightly (the body itself: not the spin axis) as it maintains its moment of inertia constant around this arbitrary axis. This produces what, under ideal conditions of no dissipation of energy, is a constant periodic motion of the body around the spin axis due to polhoding.[10] As the spinning body tends towards perfect mechanical (and in the GP-B case, electrical) sphericity, the polhode period tends to infinity. So, for the near-perfect GP-B gyros, the period of the polhode motion was expected to be very long. Furthermore, for GP-B with only minimal differences between the moments of inertia about the three principal (inertial) axes of the rotors, neither the polhode path nor its period was expected to change significantly over the duration of the experiment (Silbergleit et al., 2009).

With the evidence of electromagnetic interference perturbing the motion of the GP-B gyros, it was possible that the assumption of constant polhode motion was not valid. It was still unclear, however, what had gone wrong. But if the polhode motion—defined by the polhode angle and period—had not been effectively constant over the Science Phase of the experiment, then many assumptions might need to be re-examined and it could have far-reaching consequences. Without understanding what had gone wrong and exactly how, it seemed impossible to salvage anything meaningful from the noisy dataset with its unexpected and seemingly unpredictable jumps in rotor orientation.

It seems clear now that the "'un-modeled' forces between the gyro rotor and its housing" (as quoted above on page 147) that meant it was actually impossible to fly the spacecraft in its primary drag-free configuration around one of

[10]The polhode motion of the Earth, also called its free nutation, is normally considered to be a "wandering" of true north (south) across the surface of the globe (by about 9 m with a period of 433 days). However, this is a consequence of our taking as our reference a fixed point on the surface of the Earth (or more accurately, a combination of references from different points, to allow for plate tectonics). If we were to adopt a perspective from inertial space, then the spin axis orientation of the Earth would remain constant and the globe would be seen to wobble around it. Indeed, it was this wobble with reference to celestial bodies that allowed the polhode motion of the Earth to be detected and measured.

the gyros acting as the proof mass were the larger than expected Newtonian torques acting constantly on all four gyros, principally at the gyro spin rate, modulated by both the spacecraft roll rate and the (changing) polhode period.

5.3 The Culprit Exposed

5.3.1 Patch effects

In the light of their findings and calculations, the GP-B team were led to a conclusion concerning what had gone wrong. If we recall the table in Chapter 3 (page 84) containing the 7 technical requirements for the success of the mission, as a result of their investigation it quickly became the belief of the GP-B team that only six of the 7 had been met. I can now complete that table with the verdict that the team arrived at concerning each requirement, as they first published it in their 2008 Science Results report (Kahn, 2008, p. 12) (Table 5.1). As I have indicated, the primary justification for this belief came from the insight gleaned during the post-Science Phase calibration manoeuvres; but it also came from the further considerations concerning what could have caused electromagnetic interference which would have manifested itself in the SQUID readout signal. The team followed various leads and considered different models that would yield the type of interference they had detected; one of the most revealing of which I will now consider briefly.

As part of the GP-B team's efforts, Keiser, Kolodziejczak, and Silbergleit, 2009, derived the torque that would result from an uneven electromagnetic flux distribution on a spinning rotor in extremely close proximity to another electrostatic potential. In the GP-B case, this latter potential would result from the uneven charge distribution on the different components that make up the gyro housing and that were only microns—on average, the gap was just $32\,\mu\mathrm{m}$!—from the rotors themselves. The housing contained the gyro suspension electrodes and the SQUID pick-up loop. Patches on the surface of the rotors would cause the interference that was seen as noise in the SQUID readout as they spun very close past the pick-up loop, as a more sinusoidal interference if they were closer to the "poles" (the spin axis) of the gyro; or as

Rotor Properties		
Density homogeneity	$< 6 \cdot 10^{7}$	met
Sphericity	< 10 nm	met
Electric dipole moment	< 0.1 V·m	issue
Environment		
Cross track acceleration	$< 10^{-11}$ g	met
Gas pressure	$< 10^{-12}$ torr	met
Magnetic field	$< 10^{-6}$ gauss	met
Mixed		
Rotor electric charge	$< 10^{8}$ electrons	met

TABLE 5.1: Seven "near-zero" GP-B parameters required for a successful mission. Six of them were clearly achieved; the failure to meet the remaining objective led to unwanted interference perturbing the expected science signal which inevitably reduced the accuracy of the final results.
[SOURCE: Kahn, 2007]

a more step-like interference if they were closer to the "equator" of the gyro. Moreover, if there was an uneven electromagnetic potential in the vicinity of the gyros, they would have experienced a varying torque as they interacted with it. Such unevenness was expected on the gyro housing, as it contained different electrical elements used for the initial suspension of the gyros and the later detection and adjustment of their positions; and also the channel that acted as the helium nozzle to initially spin up the gyros: see Figures 5.5 and 5.6. When referring to this torque, the authors explain that:

> The magnitude of the torque depends on the magnitude and distribution of the patch effect potentials on the surface of the rotor and the housing, but it is important to note that this torque is due to the interaction of the patch effect on the rotor with the patch effect on the housing. Neither surface alone will produce a torque.
> (Keiser, Kolodziejczak, and Silbergleit, 2009, p. 389)

Using spherical coordinates and taking the principal axis as pointing in the (apparent) direction to the guide star—that is, the spacecraft spin axis orientation during the GSV period—they show that the resultant torque is proportional

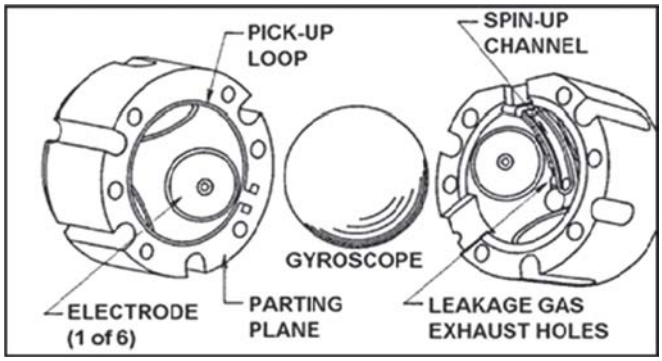

FIGURE 5.5: Sketch of the main components of the two separate halves of the gyroscope housing. The housing contained a total of 6 electrodes to guide and maintain the gyroscope at its centre; and also to sense the position of the gyro when it was acting as the proof mass in drag-free flight. The housing also contained the channel through which helium was injected to spin up the gyro during the IOC phase and the holes to evacuate the gas afterwards. The pick-up loop connected to the SQUID was located at the join between the two halves of the gyro housing.
[SOURCE: Buchman et al., 2015]

FIGURE 5.6: Photograph of an actual GP-B gyro and its housing.
[SOURCE: Everitt et al., 2009]

to the misalignment for small angles only. Since the SQUID output data generated during the calibration phase also showed a torque that was proportional to misalignment for small angles but not for larger angles, they treat this as prediction agreeing with observations (Keiser, Kolodziejczak, and Silbergleit, 2009, p. 388) (though clearly it would not stand up to criticism of being an example of non-novel use of data to derive a hypothesis that the data are already known to agree with). Since it may be argued that a small-angle approximation to linearity is a common mathematical feature and could result from several different effects, it hardly seems justified to treat such a common characteristic as proof of their specific conclusion that is was uneven electromagnetic potential distribution (later determined to be due to the formation of a dipole layer on the rotor surface) that caused the interference. Taken on its own, such agreement is in no way a stringent test of their hypothesis: it fails to rule out other possible causes. Be that as it may, their finding must be seen as agreeing with the team's final overall verdict and therefore as evidence in its favour. As in most science, it is only in combination with other evidence that this can be taken as part of the justification for the team's final judgement.

In a similar way, the team's other investigations all led to the belief that the anomalies that were detected in the GP-B data were the result of contact potential differences (that is, sustained potential differences over parts of a conductor) between different regions of the rotor surfaces which were unexpected and interacted with the (expected) potential differences on the gyro housing. This means that for some reason, electromagnetic potential—which can be thought of equivalently as magnetic flux[11]—had formed (due to some build up of unevenly distributed charge) on the superconducting rotor surfaces and the resultant electromagnetic field was interacting with that present in the immediate environment of the gyros, thus producing both interference

[11]The patches that cause these effects can be viewed in different ways. Macroscopically it is uneven charge distribution on the superconducting surface of the gyros, or the corresponding trapped magnetic flux "frozen" onto the superconducting surface. However, with the tiny effects we are dealing with here, at the possibly more appropriate microscopic scale we have magnetic dipoles formed of stationary fluxons effectively embedded and stationary within the superconducting niobium surface of the spinning gyros. These fluxons can in turn be seen as pairs of N and S magnetic poles each of one quantum of magnetic flux, connected by a vortex passing through the fused quartz body of the gyro.

in the SQUID output signal and a classical torque on the gyros; both of which varied as the gyros spun. We are told in the Science Results report (Kahn, 2008, p. 12) that in his 1989 PhD thesis, T. W. Darling describes patch effects, as he names the phenomenon, or contact potential differences that can be caused on metal surfaces at low temperatures due to variations in the crystalline structure of the surface that can result in a dipole layer forming. Prior to the GP-B data collection and analysis, it was believed that the crystal structure of the rotor surfaces was so fine that it would not give rise to any such patch effects. However, when the anomalies in SQUID readings were detected and the electromagnetic nature of the interference causing them became apparent, this assumption was re-examined. The team set about using the redundant, back-up or leftover flight-quality rotors that they had available to them on the ground to study the possibility of these so-called patch effects being the cause of (at least one component of) the interference that had been detected. Detailed UV scanning measurements revealed that their confidence in the assumption of a near-zero electric dipole (no patch effects) had been unfounded.

In their overview of the data processing published in 2009, Everitt et al. seem to suggest that the problems encountered in the readout data, at this stage—in light of all the evidence—believed by the team to be due to trapped magnetic flux within the gyros themselves, were the result of the rotors initially being cooled in a finite electromagnetic field (Everitt et al., 2009, p. 56). It would seem clear that this is where any such remnant flux did indeed come from as it was in fact always known that there would be some tiny amount of trapped magnetic flux, since reducing the magnetic field around the rotors to zero was impossible: it was one of the *near* zeroes. The expectation, however, was that the even distribution of this magnetic flux (or electric potential) over the near-perfect superconducting surface of the rotors would not give rise to a magnetic moment that could interfere with the signal from the LM; as was now believed to have actually happened, due to the formation of a dipole layer. Near-perfect electrical sphericity of the rapidly spinning gyros would have meant that the effects of any trapped flux on them would have cancelled out and would have caused minimal noise and no net torque. In fact, it seems to have been unforeseen imperfections in the surface of the rotors themselves that arose from the coating process and edge effects left in the uneven final

layers of the niobium coating that led to the problems and the "deeper than expected" (Everitt et al., 2009, p. 53: another great GP-B team understatement) data analysis that proved necessary for the team to recover the relativity signals.

5.3.2 Engineering data to the rescue

Even when patch effects had been identified as a source of the interference with the LM signal detected by the SQUID electronics, and also as producing larger than expected Newtonian torques on the gyros, it was unclear how meaningful results could be salvaged from the noisy data. It was also still unclear what had caused the occasional jumps in gyro spin axis orientation. Once again, however, the meticulous experimental design, allowing for as many unforeseen circumstances to be dealt with as possible, was to prove invaluable. Fortunately, the engineering requirements (which at one and the same time represent part of the built-in redundancy of the entire mission) meant that in addition to the continuous collection of science data, short bursts or snapshots of far more detailed engineering data were also regularly collected.

The science data, as we know, provided the datapoints that were plotted to show how the co-moving telescope and spacecraft system containing the SQUID pick-up loop shifted in orientation compared to the gyro LM about which it rolled. As we have seen (3.6) and as should be very clear from the experimental set-up, every 77.5 seconds, as the spacecraft completed one roll around the gyro, the SQUID voltage output completed a sine wave as the LM moved across the pick-up loop and (virtually: the minute GTR drift was also present) back to its staring position. To record this, the output voltage was sampled or measured twice per second;[12] this is the recorded science data

[12]In fact, the SQUID readings were taken at the slightly higher frequency of 5 Hz: five times per second, and this constituted the L1 data. These data were then pretreated to arrive at the L2 datapoints which were two every second. This allowed for outliers to be removed, and all on-board systems to be synchronised as corrections were necessary due to the specifics of the on-board circuitry. Silbergleit et al. tell us that in this way "a single L2 data point at a predetermined moment of vehicle time

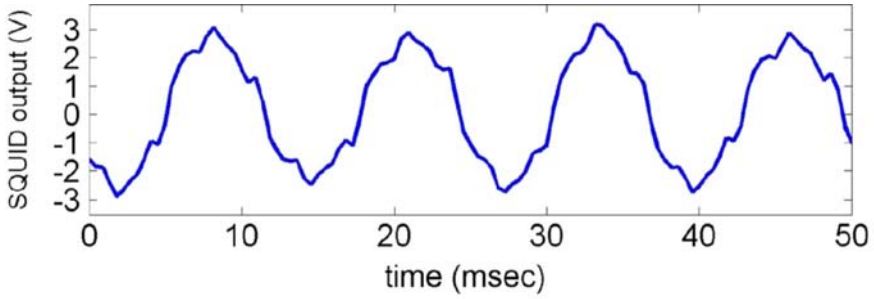

FIGURE 5.7: The figure shows the GP-B SQUID output engineering data, sampled 2,200 times per second. At this frequency, the wave form of the interference is clearly visible in the signal. To understand the problem for the science data, we have to imagine the data collected here over a 2 second period condensed into just one value and then added to the science data. Clearly the interference at that frequency appears as white noise that cannot be modelled.
[SOURCE: Silbergleit et al., 2015b]

(which is converted to an angle via the conversion factor established for the system, as I have mentioned previously). This is a low-frequency (LF) signal, since it is collected just twice every second: a 2 Hz signal (therefore each sine wave consists of 155 datapoints). The far more detailed bursts of engineering data, in contrast, sampled the SQUID readout voltage at a frequency of 2.2 kHz: 2,200 times per second. Each such high-frequency (HF) snapshot only lasted 2 seconds, but one was recorded approximately twice every minute, when the on-board conditions and systems permitted (Silbergleit et al., 2015a, p. 14).[13] Recall, the gyros were spinning at frequencies of around 70 Hz (from 62 to 82 Hz), so the 30-odd engineering datapoints per complete gyro rotation or spin, continuously recorded over some 160 consecutive spins (in each 2-second snapshot) is what allowed the GP-B team to observe the variations from smooth sinusoidal science data in detail.

These detailed snapshots of the SQUID output voltage readings were thus the

was derived from a number of neighboring L1 data points" (Silbergleit et al., 2015b, p. 10)

[13]We are told in the paper that the longest gap between snapshots during the entire mission was as much as 2 days; but this was clearly highly exceptional as nearly 1 million (976,478) were recorded in total and successfully processed afterwards.

key to confirming the team's suspicions, as they allowed them to "see" within each burst of this detailed information, the details of the perturbations in the readout, which appeared simply as white noise in the LF signal. Through examination of the HF engineering data, it was possible to observe the regular perturbations in the readout at the gyro spin frequency: see Figure 5.7. This confirmed that there was highly exaggerated electromagnetic interference with the SQUID readout signal at the spin frequency of each gyro which could not have been produced in the absence of an uneven electromagnetic potential distribution on the gyro surface. The effect was detected for each of the four gyros; each with its individual spin rate. By comparing subsequent snapshots it also became apparent that the interference was modulated at the polhode frequency of each gyro, as the patches effectively moved around relative to the gyro spin axis.[14]

As I explain above, (starting on page 154), polhode motion, or polhoding, results from the tiny deviation from exact symmetry of a spinning body. It is effectively the slight wobble of the body around its axis of rotation since its mass is not perfectly evenly distributed about that spin axis. The period of the wobble can be measured and for torque-free spin should be constant. (The effect was initially discovered and measured for the earth in the 19th century.) Since the design requirements for the symmetry (and homogeneity) of the gyroscopes was another *near* zero, there would inevitably be some residual polhoding of the gyros. Those near zero limits placed on the deviation from sphericity and perfect homogeneity, coupled with the rigidity of the fused quartz the gyros were made of, supposedly ensured there was (virtually) no energy dissipation from the gyros as they spun in their near vacuum. Under these conditions, the polhode motion was expected to be constant, (except for a slight dampening of the polhode period after initial spin up, as the gyro settled into its steady state).

Exact knowledge of the phase of this wobble, or polhode phase, was always necessary for the GP-B calculations, but it was believed that with an effectively stable polhode path and period, this effect could easily be built into the

[14]Of course, the patches were actually stationary on the surface of each gyro and it was the entire gyro that was moving relative to its spin axis: this is the gyro polhode motion.

calculations. The polhode period and phase formed part of the calibration of the entire science signal output, and the constant period of this motion was built into the way in which each partial GSV dataset was added to the previous and next, to form a continuous data stream. On each orbit around the earth as the GP-B spacecraft came out of occlusion and the telescope acquired the guide star again, the exact polhode phase was used in the calculations to connect the data from the new guide star valid (GSV) period with that from the previous GSV period. In this way it was possible to put together all the (approximately 60-minute long) GSV readings into one continuous stream of data.

Technically, the polhode phase is used in the calculation of **G**, the gyro scale factor, which is performed using the SQUID readout and is necessary to determine the absolute angle between the telescope (or equivalently, the spacecraft) pointing direction and the gyro spin axis orientation. (Although without a precise determination of **G** the relative change in the angle between the two directions could be calculated, to acquire an absolute measure of this angle—and thereby to obtain the desired relativity data—requires exact calibration which includes allowing for the polhode angle in the direction given by the polhode phase in the "dirty" signal or level 1 (L1) output data.) The fused quartz that the body of each GP-B gyro consisted of was rigid enough and their sphericity and homogeneity sufficiently near perfect for there to have been effectively no energy dissipation under torque-free spinning of the rotors. Such a situation—that expected for the GP-B mission—without the anomalies that were attributed to patch effects, would have meant that the effectively constant polhode period would have made the polhode motion easy to allow for.

Any significant torque applied to the gyro would change that situation. Moreover, if the polhode period was changing during each orbit, it would be impossible to know the exact polhode phase when the guide star was reacquired and effectively the gyro would have to be recalibrated on each orbit. Even for the fused quartz the gyros were made of, external torques will cause stresses and strains on the crystalline structure which will inevitably result in the dissipation of tiny amounts of kinetic energy and heating of the crystalline

structure. Consequently, the polhode motion will vary as the rotor body turns, adopting the position in which, for a given (conserved) angular moment, it has the least possible kinetic energy.[15] This minimum of kinetic energy occurs when the spin axis of a rotor coincides with the axis through the centre of mass of the rotor with the largest moment of inertia: its major principal axis. In this way, while the spin axis remains fixed in inertial space, the rotor body moves around it, so to speak, until it is moving around the principal axis of the rotor with the greatest moment of inertia. This appears to be exactly the evolution that all four GP-B gyros underwent during the mission. So, since constant polhode motion was precisely what it was believed had not in fact been the case, it became vital to track the polhode path and period in order to devise a revised analysis of the GP-B gyros.

5.3.3 Effects of effects . . . of effects

So, the engineering data enabled the team to establish that there was interference in the science data (SQUID output voltage signal) at the spin frequency of the gyros and that this was modulated at their polhode frequencies. The evidence from the work carried out on the flight-ready gyros on Earth, indicated that this was due to the formation of patch effects that would account for the electromagnetic nature of the anomalies detected in the post-science calibration. Moreover, the torques exerted on the gyros by the trapped magnetic flux as it rotated through the electromagnetic fields associated with the gyro housing led to a tiny degree of energy dissipation and thus a variation in the polhode frequency, which was also observed in the engineering data. There was still no explanation for the seemingly random jumps that were occasional observed in the gyro spin axis orientation. The full solution to this

[15] In the Science Results report it is noted that the physical basis for this energy dissipation is not clearly understood. This is a point I will return to when criticising the explanation given; assumptions such as "stationary" fluxons are made in order to perform the TFM that I discuss in the next section, while the team members recognise that they do not know—or have much idea—what is actually going on in a rigid spinning body dissipating energy and suggest that maybe inelastic forces within the body of the rotor play an important role but "remarkably" do not affect their ability to map the trapped flux accurately (or, apparently, their confidence in the method).

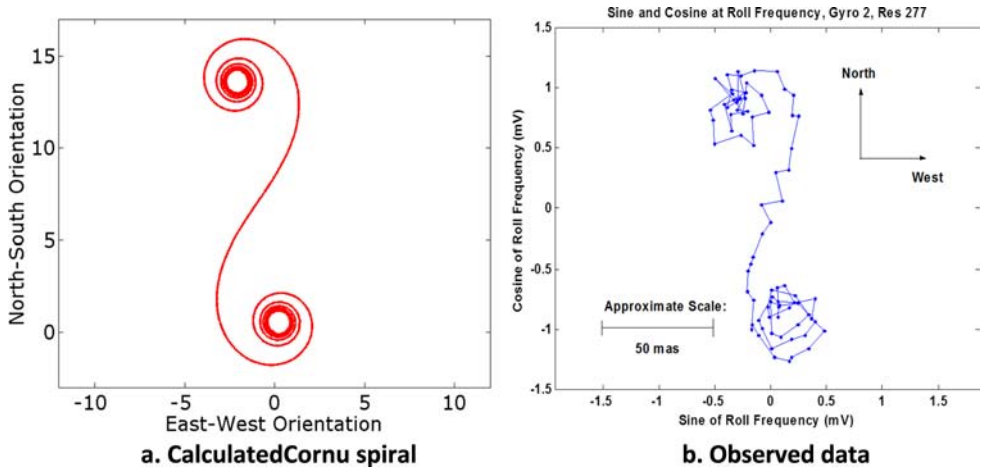

FIGURE 5.8: On the left, we can see a plot of an Euler (or a Cornu) spiral. It is defined as joining a circle to a straight line by following a path whose curvature is proportional to the length of the curve. It is the curve that a resonance boost would be expected to produce in an oscillator as the two resonances come together to coincide and then move apart again. On the right we see the actual path plotted for one of the jumps in rotor orientation recorded for a GP-B gyro. In the GP-B team's words the Euler spiral is "exactly what was observed".
[SOURCE: Everitt et al., 2015]

mystery would not become clear until a method had been devised to clean up the SQUID output signals by modelling the interference that the signals suffered due to the patch effects. I will explain that solution in the next section, but here I just want to advance the resolution of this issue.

When the jumps were originally plotted (as we saw in Figure 5.2 above), the gyro axis orientation was shown to undergo a characteristic spiralling out, then a translation and finally a spiralling back in to the new orientation. This type of motion is immediately reminiscent of a resonance effect. However, there was initially no indication of what two frequencies could be entering into resonance with each other to magnify a force on the rotor and cause it to jump in this fashion. Figure 5.8 shows the calculated theoretical spiral motion that is caused by a resonance effect providing a boost to an oscillator and diverting it from its original position as the resonance frequency is approached and then passed; this is the Euler (or Cornu) spiral. It is compared, on the right, to the

characteristic jump of a GP-B gyro we have already seen (Figure 5.2). Analysis of the polhode motion and its changing frequency as the mission advanced (due to the poorly understood tiny degree of energy dissipation) revealed that these large individual excursions occurred when there was resonance between a high-order harmonic of the polhode period and the steady roll rate of the spacecraft. Working always to lowest order (l.o.; as justified in their paper) we are told that:

> the gyro spin axis during a resonance makes a "step-over" in the NS–WE plane following, to l.o., a Cornu spiral winding out from its initial direction before the resonance, moving across, then winding back in to the new direction.
>
> (Silbergleit et al., 2015a, p. 48)

As I say, the issue was not settled until the modelling was quite advanced, as it depended on the calculation of the polhode frequency throughout the entire history of the Science Phase. However, we now know that we have an interference with the output signal at the gyro spin frequency modulated by the polhode frequency; and it was now expected that the latter would be changing due to energy dissipation. The full model that I describe in the next section took as its starting point electrostatic patch effects on the gyro rotors, causing HF interference in the SQUID output signal. The interference was shown to be at the spin frequency of each gyro modulated by the polhode frequency, and it confirmed that the jumps coincided with the occasions when a harmonic of the polhode frequency of the gyro coincided with the spacecraft role frequency, as the polhode frequency changed over the lifetime of the experiment.

The match between the actual data, on the right-hand side of the figure (Fig. 5.8), and the theoretical prediction allowed Buchman et al. in another of the Classical and Quantum Gravity Focus Issue papers to affirm that:

> offsets in the orientation of the gyroscope spin axis occurred when a high harmonic of the gyroscope polhode frequency was equal to the satellite roll frequency. A reexamination of the patch-effect torque formalism, showed that these roll-polhode resonance torques could also be explained by the same formalism used to analyze the misalignment torques when no averaging over the satellite roll phase or the gyroscope polhode phase was assumed.

> Moreover, integrating over the drift rate during one of these resonances showed that the path of the gyroscope spin axis followed a Cornu spiral, which was exactly what was observed.
>
> (Buchman et al., 2015, pp. 22-3)

Let me just summarise the three effects that were now believed to have interfered with the science data, all resulting from the patch effects.

1. Electromagnetic interference in the SQUID output signal. This was due to the patches of flux frozen in the gyro shell (or the contact potential differences formed on the shell surface) being detected by the SQUID as they passed by the pick-up loop (twice) with every spin of the gyro. It was modulated at the polhode frequency as the gyro body turned itself to the orientation of minimum energy while maintaining its angular momentum constant.

2. Unexpected (or unexpectedly large) classical (Newtonian) torques on the gyros as the patch effects on them interacted with the electromagnetic field generated within the gyro housing. Again, this was at the gyro spin frequency[16], but just as with the interference, modulated at the (changing) polhode frequency, as the patches moved their way (relative to the housing and pick-up loop) over the surface of the gyro.[17]

3. Finally, those unexpected torques modulated at the (changing) polhode frequency occasionally entered into resonance with the (constant) spacecraft roll frequency. This only happened with high-order harmonics of the polhode frequency, but effectively at these frequencies the torques were amplified and pushed the gyro off its axis, resulting in the jumps observed in gyro spin axis orientation.

It is often said that knowing what the problem is, is half the battle of solving it. That may be optimistic, but armed with the knowledge they had derived from their preliminary investigations, the GP-B team set about devising a method to calculate and remove the systematic errors introduced into their unique

[16] This is actually the frequency relative to the housing, as with the interference, so it is at gyro spin frequency $\pm$ spacecraft roll frequency.

[17] Of course, it is the whole gyro that is twisting and the flux that remains frozen in the gyro surface.

dataset by the effects of the unexpected patch effects (and the secondary effects of those effects of the patch effects). In fact, it was in 2008 that the team affirmed:

> These patch effect terms are now known to explain the two classes of anomalous Newtonian torques, and quite probably also the changing polhode period.
>
> (Kahn, 2008, p. 12)

The method that they devised was rather prosaically called trapped flux mapping (TFM): it is just that, and is what I consider in the following section.

5.4 Trapped Flux Mapping

5.4.1 The objective of TFM

It would be hard to exaggerate the lengths that the GP-B team went to try to salvage reliable results from their noisy dataset. If we first consider the time spent on the data analysis, we get some idea: from less than one year, as originally intended pre-launch, to over five and a half years. The financial cost may be another indication; though more difficult to establish (although much of the data may be in the public realm, it is not publicised; also there may be undisclosed sums coming from different sources and we face the perennial difficulties involved in calculating proportions of time spent on different projects and the cost of support services, and so on). To get a very rough idea though, we can take the figures we are given for the stopgap funding solution which ran from 15/01/2008 to 30/09/2008: a period when the process was well underway, but still had a long way to go. In those eight and a half months, we are told that the funding was a total of US$ 1.715 M (Kahn, 2008, p. 1): scaling that up to the total length of time of the data analysis would give a (very crude) estimate in the region of US$ 14 M as the data processing budget. So it seems that eventually, despite NASA's withdrawal from the project, neither time nor funds were lacking. In order to see how the process was developed, let me just recap some of the basics.

When a type-II superconductor—such as the niobium shell of the GP-B gyros—is cooled below its critical temperature, any residual magnetic flux is frozen or trapped in the superconductor. As I have said, this trapped magnetic flux takes the form of pairs of magnetic quanta spatially separated within the superconductor and effectively connected via a magnetic vortex line, which, for the shells of the GP-B rotors, passes through the body of the rotor itself. These magnetic quanta have an associated magnetic field which is registered by the SQUID pick-up loop as it rotates about the rotors at the spacecraft roll rate, but also as the rotors themselves spin. This is not the case for the LM magnetic field, which points along the rotor spin axis orientation and therefore is only registered as part of the low-frequency (LF) signal resulting from the spacecraft roll. This difference is crucial to the analysis of the SQUID signal as it means that the two different sources produce signals with completely different, and easily identifiable, signatures. In the abstract to their paper addressing trapped flux and GP-B data analysis, Silbergleit et al. mention that the HF signal is at the gyro spin frequency.[18] They also tell us in the Introduction that the HF signal is a "superposition of multiple harmonics of the spin frequency" (Silbergleit et al., 2009, p. 397).The important point for identification and treatment of the signals is that a typical GP-B gyro spins at between 3,700 and 5,000 rpm ($f \sim 100$ Hz) whereas the spacecraft rolls every 77.5 s ($f \sim 0.01$ Hz).

This difference in signatures, with the two components of the total SQUID output signal (that from the LM and that from the trapped flux) producing signals at their respective characteristic frequencies, was the crucial feature that allowed the team to even consider disentangling them and recovering the GTR signal. Because of these widely different signatures, the relativity data were still expected to be found in the LF signal, but in a more complicated fashion as the changing polhode period (combined with the Newtonian torques exerted on the rotors) meant that the polhode motion had to be modelled for each gyro in such a way as to combine it with a stationary map of the flux trapped on the surface of the gyro. The aim was thus to model the

[18] As I have already indicated, it is in fact at gyro spin frequency $\pm$ spacecraft roll frequency, but as the spin rate is so much faster than the roll rate, the approximation in their explanation (not the mathematics!) is quite natural.

trapped flux as it moved, spinning at the gyro spin frequency but in a fashion modulated by the polhode frequency; while this latter was itself changing due to the effects of the trapped flux interfering with the electromagnetic field in the neighbourhood of the spinning gyro (produced by the housing). Clearly this was a highly non-linear process which required complex modelling and a reiterative process to hone the models.

So, the two key elements in the approach adopted by the team to resolve the entire issue were a map (TFM) of the patch effects frozen on the gyro surface and the accurate determination of the polhode angle and phase throughout the entire duration of the mission; both individually for each of the gyros, of course. As they describe it, the aim was to map the trapped flux onto the surface on each gyro to provide a model of the source of the modulation of the polhode motion observed in each gyro.[19] As was implied above, the varying polhode period initially seemingly made it impossible to calculate the SQUID scale factor for the gyros. However, now, if the team could track the (unexpectedly changing) polhode period, and its phase, they had a chance of separating out the two components of the SQUID output signal. They went on to develop, not one but two separate methods to determine the precise configuration of the patches that was causing the problems. For now, let me just quote Buchman et al. from their contribution to the Classical and Quantum Gravity Focus Issue in which they make it all sound so straightforward:

> The observations and analytical confirmation of these two torques due to the patch effect—the misalignment torque and the roll-polhode resonance torque—allowed these effects to be included in the GP-B data analysis and separated from the relativistic drift rates.
>
> (Buchman et al., 2015, p. 23)

[19] I hope the ironic analogy with GTR does not escape the reader. In that theory, it is the precise distribution of energy–momentum that determines the topology of spacetime; while the topology of spacetime is precisely what determines how energy–momentum behaves. Here, it is the trapped flux distribution that determines how the polhode motion changes; and it is the changing polhode motion that provides us with the clues as to the trapped flux distribution.

5.4.2 How TFM works

When analysing polhoding, an important parameter is the inertial asymmetry parameter, Q, given by:

$$0 \leq Q \equiv \frac{I_2 - I_1}{I_3 - I_1} \leq 1$$

where the (increasing) subscripts denote increasing[20] values of the principal moments of inertia. Despite the extremely precise sphericity of the GP-B rotors, for which the principal moments of inertia obey: $(I_i - I_j)/I_k \sim 10^{-6}$ (this expresses the degree of sphericity, by representing departure from perfection), Q relates the asymmetry in two different planes, and therefore can take a value anywhere between 0 and 1 for any imperfect sphere (however slight the imperfection). Q takes the value of 0 for any (near) sphere whose moments of inertia about the two orthogonal axes both of which are orthogonal to the axis with the greatest moment of inertia (principal inertial axis) are equal; that is, $I_1 = I_2$, which is fulfilled for a body that is symmetrical about its axis of maximum moment of inertia (that is, for an oblate sphere, a torus, a flat square and a disk; but not a long cylinder). The values of Q for the GP-B gyros (as estimated from the TFM) varied from 0.13 to 0.30 (Silbergleit et al., 2009, p. 405). However, it is important to note that in the initially assumed case of stable polhode motion, this factor would have been totally irrelevant, and in fact even a value tending towards 1 (for example, that of a disk) would have had no effect on the relativity measurements. This factor is only important because it affects how the gyros turn on themselves in order to minimise their kinetic energy while preserving angular momentum. This energy dissipation is presumably caused by the interaction of the electrostatic patches on the rotor surface and the electromagnetic field in the gyro housing, as this was postulated as the root cause of all the perturbing effects.

As I indicate above, the total signal (or flux through the pick-up loop, which is proportional to the signal) produced in the SQUID pick-up loop can be separated into two components. The first of these (and the only one originally intended to make a significant contribution) is produced by the LM. This is

[20] Technically, as pointed out in the paper, this is the "non-decreasing" order as two values (or all three) could be equal.

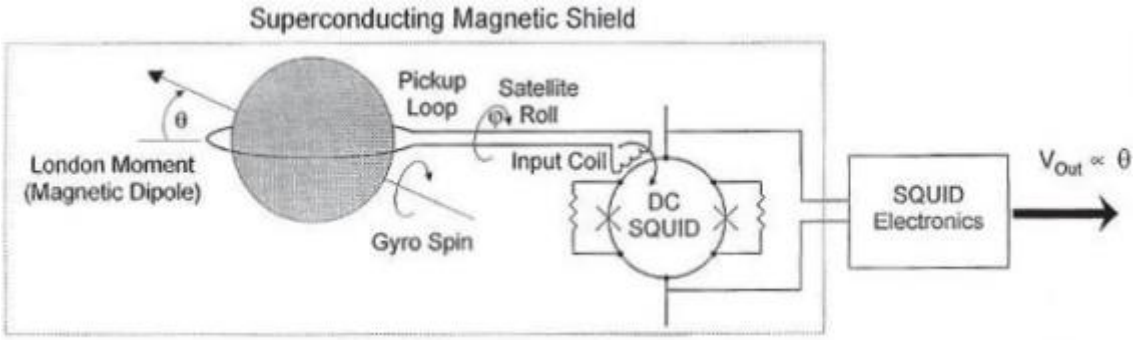

FIGURE 5.9: Representation of the London moment that is generated by a spinning superconductor and aligned precisely with its spin axis. This is the effect that was used to determine the orientation of the gyro spin axis in GP-B.
[SOURCE: Muhlfelder et al., 2015]

in the direction of the gyro spin axis which, due to the specific experimental design, is very nearly in the plane of the SQUID pick-up loop and therefore it is proportional to the tiny angle between the two, θ (see Figure 5.9). (However, as the magnetic moment produced by the gyro moves further from the plane of the pick-up loop, this proportionality becomes less perfect and should be substituted for $\cos\theta$; as used in TFM.) This is precisely what the experiment was designed to measure: the spin axis orientation relative to the telescope (spacecraft) pointing direction. To obtain the absolute spin axis orientation, the SQUID signal (proportional to the pointing angle, as I have just said) needed to be very accurately calibrated.

For the LM, lying very nearly in the plane of the pick-up loop[21], the scale factor is proportional to the gyro spin speed and modulated only at the spacecraft roll rate (and therefore constant to the spin-down rate of the gyros). The second component is the (undesired) interference from the magnetic fluxons trapped in the gyro (the patch effect on the surface of the gyros) that are spinning and wobbling. The fluxons (or electromagnetic potential) in the

[21]Orbital aberration is ~ 5 seconds of arc and annual aberration is ~ 20 seconds of arc which together with the predicted geodetic effect could lead to a (worst case) maximum misalignment of $\sim 0.01^{\circ}$; assuming an actual cosine dependence, the linearity approximation is still good to $1 in 10^7$ at this misalignment. To approach the 1:1,000 level, the misalignment angle has to be $\sim 2^{\circ}$.

surface of the gyros are passing extremely close to the pick-up loop as the rotor spins. As I mention above, unless they are very close to the "pole" of the gyro, the signal they induce in the SQUID is virtually a step-function as they approach the pick-up loop at the spin rate $\pm$ the spacecraft roll rate (recall that 2 of the gyros are spinning clockwise and two anti-clockwise; therefore, for the 2 that are spinning in the same sense as that in which the spacecraft is rolling, the rate of approach is $\phi_s - \phi_r$, whereas for the two that are spinning in the opposite direction it is $\phi_s + \phi_r$), pass within $32\,\mu\mathrm{m}$ of the pick-up loop and then recede at the same rate. This motion of the trapped flux is modulated by polhoding: as the gyro polhodes, so this component of the signal (almost stepped HF, trapped fluxon component consisting of multiple harmonics of the (spin $\pm$ roll) rate) varies according to the (time-varying) polhode motion of the rotor. We can best understand this by considering the Figures 5.10 and 5.11. The former shows the LF output that was expected from the SQUID, with the 2 Hz signal tracing out a sine wave very 77.5 seconds as the spacecraft rolls around the LM, which is constant with respect to the gyro spin frequency. The latter, shows the HF signal produced by the total flux trapped in one of the gyros as it spins with the gyro.

From FFT analysis of the (first 6) harmonics of the spin frequency, the GP-B team claim to have been able to decipher "the full time history of the polhode period for each of the GP-B gyros". This can be gleaned from the (very gappy) batches of SQUID engineering readout data, taken at 2,200 Hz with 2 s stretches of data being available every 40 s (on average). Furthermore, they claim that the polhode history was visible in the modulation of gyro spin in the LF signal that was the constantly monitored science signal. These different sources of polhode period history agree very well for each of the gyros, while they individually behave very differently (as we should expect, considering that the random initial spin may have been around any inertial axis, but it will tend to the principal inertial axis). This agreement leads to the confidence in this being the actual polhoding of the gyros.

Confident that they now knew how the gyros were moving, the difficulty lay in explaining why they were moving in this unexpected fashion. It was

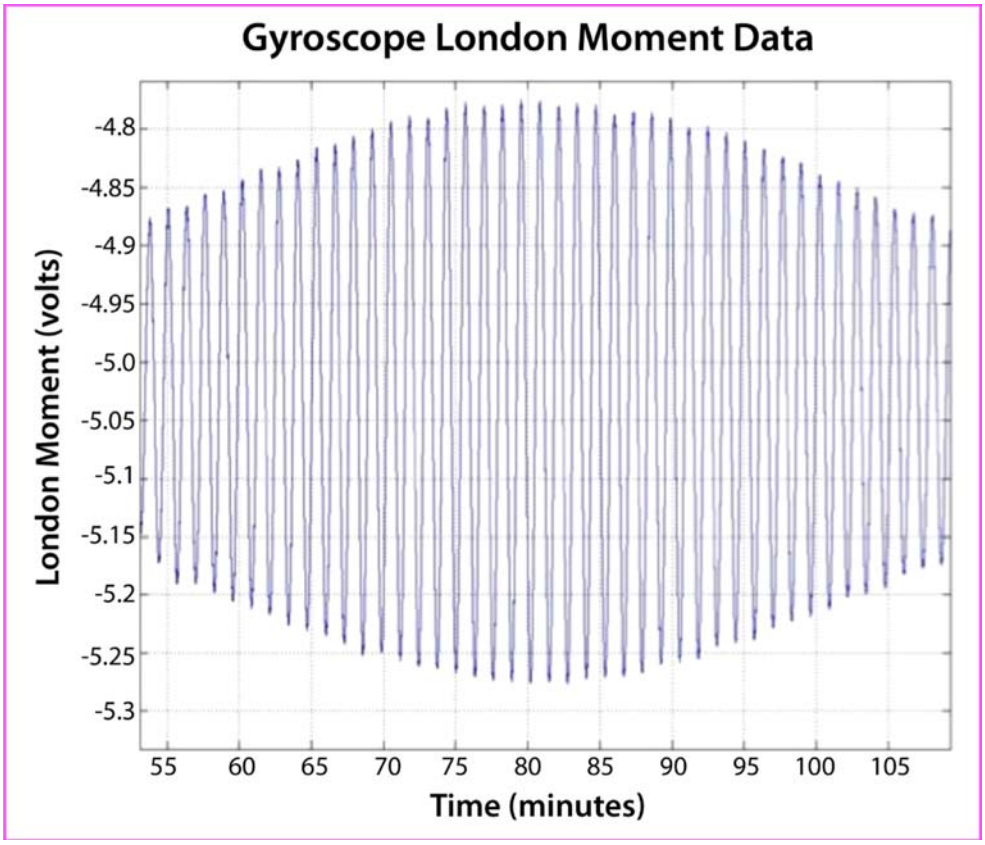

FIGURE 5.10: The figure shows the expected GP-B readout science data. As the SQUID pick-up loop rolls around the gyro, the gyro LM produces a voltage signal in the SQUID circuitry. This signal varies at the roll rate of the space craft (completing one full sine wave every 77.5 seconds, as can be seen). [SOURCE: Everitt et al., 2015]

attributed, as I said above, to energy dissipation. Using the maximum asphericity for the gyros given above, Silbergleit et al. calculate the total energy loss—that required to turn the gyro from spinning about its minimum inertial axis to spinning about the maximum inertia axis—as $4\,\mu J$; an average dissipation rate over a year of 10^{-13} W. However, as I have already indicated, they themselves recognised that:

> The physical origin of this energy loss is not completely clear. It . . . probably is dominated by dissipative patch effect torques.
> (Silbergleit et al., 2009, p. 402)

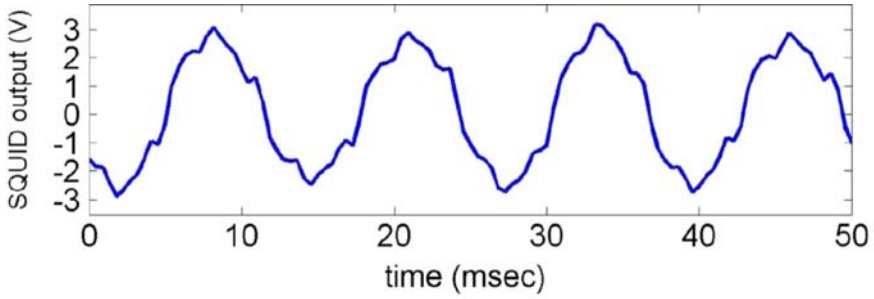

FIGURE 5.11: The figure shows the GP-B SQUID output engineering data, sampled 2,200 times per second. At this frequency, the wave form of the interference is clearly visible in the signal. To understand the problem for the science data, we have to imagine the data collected here over a 2 second period condensed into just one value and then added to the science data. Clearly the interference at that frequency appears as white noise that cannot be modelled.
[SOURCE: Silbergleit et al., 2015b]

So, having effectively characterised the polhode motion at one time, the team proceeded to derive the parameters that govern how that motion will change over time: Q; the asymptotic period that the polhoding tends to; and the characteristic dissipation time. These parameters allowed them to conclude that despite the tiny energy loss, the polhode motion was "quasi-adiabatic". This may seem strange seeing as the change in polhode motion is supposedly driven by energy dissipation! However, it is the scale of the energy loss that allows the team to justify the assumption that this process is governed by the sphericity and non-electromagnetic properties of the gyros.

With the vastly different signatures of the two components established, initially a so-called geometric method for separating the effects was devised over a two-year period starting in August 2006, and the first results were available in August 2008; 3 months after NASA had decided to halt funding. (It is interesting to speculate as to whether these results would have affected the NASA decision had they been achieved 6 months earlier.) The results were not as accurate as required, but they paved the way for further refinement and the implementation of the next stage, which continued into 2011 and produced the far more accurate results that were always the primary goal of GP-B. Once

more computing power was made available, the team was able to implement a more complex method to refine their results and indeed to compare the results of the two separate methods. (The GP-B team seems to be in no doubt as to the validity of the method and its potential for producing valid final results at the required level of accuracy, despite the doubts raised by the NASA Senior Review.) The December 2008 report from the GP-B team claims that the limit on their interim results was due to restrictions in computational power and that the "intrinsic limit of the gyro readout" would be approached once the new, high-speed computing techniques, under development at the time, were implemented.

Once again Everitt et al. inform us that:

> Remarkably, it has proved possible to provide detailed understanding of all the more important disturbing effects, and provide Newtonian methods for rigorously removing them.
>
> (Kahn, 2008, p. 11)

Remarkable indeed; but were the team justified in such confidence or were they being over optimistic as the NASA Senior Review Committee judged? With respect to the use of two different methods to determine the TFM for each gyro, which allowed the team to go on and separate out the target GTR effects, in the long-awaited publication of the final results in 2011, Everitt and his team again inform us that:

> Two data analysis methods were used to determine and cross-check the relativity results. One, called "algebraic", was based on a parameter estimator utilizing the gyro dynamics and measurement models detailed below. . . . The other method, called "geometric" . . . neatly eliminated the need to model the misalignment torque. . . . Further cross-checks from the geometric method included relativity estimates for 2 of the 4 gyroscopes, in both NS and WE directions, in statistical agreement with the algebraic results.
>
> (Everitt et al., 2011, pp. 2, 5)

In fact, it seems appropriate here to quote Everitt's team at some length, from the 2011 results published first on-line in arXiv and dated 17th May 2011, then in Physical Review Letters dated 31st May 2011. (It can hardly be a coincidence that it was precisely in Physical Review Letters that Schiff

published his original "Possible New Experimental Test of General Relativity Theory" back in 1960.)

> We performed several important data analysis cross-checks. First, the gyro drift rate results of Table II are confirmed by separate analyses of the segmented data. The 24 independent results from six data segments for each gyroscope are all consistent with the joint result within their confidence limits, demonstrating the internal consistency of the model. Second, the misalignment torque coefficients k determined during the calibration phase proved to be in excellent agreement with the end-of-mission values estimated by both algebraic and geometric analysis methods. No less impressive was the agreement between the time history $k(t)$ throughout the mission obtained by the two methods. As for the roll–polhode resonance torques, the gyro dynamics model of [the gyro equations of motion] predicts that during a resonance the gyroscope orientation axis approximately follows a Cornu spiral. Indeed, that is typically observed in orbit-by-orbit gyro orientations determined by both data analysis methods.
>
> (Everitt et al., 2011, p. 5)

5.5 Issue Resolved? How I See the Argument

As we saw in Table 5.1, in a wonderful example of what we might call "NASA speak" (refusing to call a spade a spade, as we say; or more precisely resisting the temptation to call what appears to be a gigantic cock-up precisely what everybody around the table thinks it is) each pre-mission requirement was either "met" or there was an "issue". What chance was there really of resolving this issue? There was only one unique science dataset to work with; although it was divided into four separate batches from the four gyros.[22] However,

[22] Although there were in fact 4 independent gyroscopes and therefore 4 separate datasets, the uneven charge distribution on the surface of each gyro bore no—or at most very little—resemblance to that of the others, as they were due to essentially random effects. Furthermore, which of its 3 principal axes each gyro's spin axis was most closely aligned to was totally random. (Clearly, we would expect an average of one third of all gyros to be spun up in such a way that their spin axis was most closely aligned to its principal axis with the largest moment of inertia;- 2 of the 4 GPB gyros were observed to be in such an initial configuration, with the other two executing the

there was also the calibration data which proved so useful in identifying the cause as electromagnetic. There was also the extremely detailed engineering "snapshots" of data which provided the 2-second bursts of high-resolution information on gyro orientation. But did the team succeed in identifying a cause which left a trace that was so different from the science data that its distinguishing features or characteristics allowed it to be modelled separately and removed? That is, was the unwanted effect not just gradually and steadily pushing the gyro out of line over the experimental period, or acting at the spacecraft roll rate?[23].

The resounding answer of the GP-B team, which appears to be rigorously supported by their results they have published, is that yes, the dataset can be used to distinguish between the use-constructed hypotheses they put forward and any other postulated explanation. The crucial point is that had any of their hypotheses been false, then they would not have been able to come up with such a match between hypotheses and data. Thus, although each of their claims is a clear case of double-counting—insofar as the same total dataset of the SQUID readouts was used to arrive at them as is offered as the evidence to support them—we can see them as being stringent and severe tests because they rule out alternative hypotheses and identify genuine difference makers. If the grounds for the scepticism cited in the NASA Senior Review report were simply that this was a case of double counting and therefore produced unsupported hypotheses (which could never be supported

jumps required to bring about such a situation and thereby minimise their kinetic energy. This may be deemed statistically inconclusive, but it can alternatively be used as further evidence of the fact that the modelling fits the observations.) Therefore in this instance, the fact that there were 4 gyros and 4 independent datasets did not help in the analysis, beyond showing that the analysis was equally satisfactory for the four independent configurations. This is in clear contrast to the situation with regard to the final GTR results, where being able to combine 4 independent datasets allows the overall error to be reduced.

[23]The relativity signal was predicted to gradually push the gyro spin axis away from its initial direction over the year; so if the perturbation was doing the same, there would be no way to separate the two signals out. Likewise, the amplitude that the LF SQUID readout displayed was modulated at the spacecraft roll rate (as the SQUID pick-up loop rotated around the spinning gyro), so if the perturbation were showing up at this frequency there would be no way to separate it out from the target signal.

as they relied entirely on the unique GP-B data), then according to a stringency requirement, that scepticism seems to have been successfully dispelled.[24] Notwithstanding, the doubts expressed to me by members of the NASA Senior Review Committee were of a very general nature, to do with the removal of large systematic errors from any noisy dataset. It is not clear that such scepticism has been answered.

So let me end this chapter with a brief summary of what might be an argument between one who we can consider to be a sceptic with regard to the findings of the GP-B team, and a convinced believer in all things GP-B. After their years of number crunching and millions of dollars spent, have the members of the team analysed the data sufficiently thoroughly to convince their detractors? Is GP-B an example of good science that has pushed forward the boundaries of experimental practice, or an outstanding example of how easy it is get things wrong and how difficult to get them right? Does the GP-B team occupy defensible ground that we should all recognise for the contribution to human knowledge that it represents, or should we be analysing this episode to learn how not to repeat mistakes of overconfidence, overspending, excessive determination and a lack of reasoned checks and balances on our collective efforts to understand the universe we live in? Did the team stray from the path of science and the generation of knowledge through experimentation, into the realm of speculation and even self-delusion; or have they demonstrated the almost limitless ways in which we can bounce back from adversity in all walks of life and the resourcefulness that makes scientific experimentation an endless path towards the growth of knowledge (especially through error)? In short, how would a gracious physics buff extol the virtues of the project and its findings to such a sceptical contentious individual?

GPB: The patch effects model predicts perturbations in the gyro response to misalignment that would be linear only over small angles; as was

[24]The scientist should not lose sight of the fact either that the final results that the team arrived at were from a very noisy dataset, compared to their initial expectations, and even if we accept all their claims concerning cleaning it up, their final results still contain an error of approximately 18% compared to the 1% that was their original target. This makes them only comparable to other tests of weak-field GTR, nothing like the massive improvement or by far the best test to date that it was hoped they would be.

observed.

SCI: *But such a cosine dependency of angle on effect is a common feature of many physical phenomena and therefore, this argument (while valid) is far from convincing.*

GPB: The change in frequency of the alternating current steering or piloting the gyro led to specific changes in the response of the gyro that were greater than expected; had the cause of the torques leading to those variations not been electromagnetic in nature, they would not have been produced simply due to alterations in the a/c frequency.

SCI: *But that does not show that it was trapped magnetic flux leading to contact potential differences on the surface of the gyro that caused such responses, even if they were electromagnetic in nature.*

GPB: Although the actual gyros were in orbit around the Earth, in a satellite that had gone silent,[25] minute examination of replicas of the rotors on Earth showed that they exhibited exactly the type of imperfection in the final layer of their niobium coating that made them extremely prone to the formation of dipole layers and thus to patch effects. Knowing that the design was prone to this fault and after seeing the evidence of a phenomenon that was electromagnetic in nature, a working model of patches of electromagnetic potential frozen on the surface that fits the observed interference in the data is clearly an accurate description of what actually happened to the experimental gyros.

SCI: *But this is just one such possible distribution of patch effects on the rotor surface, there could be many different arrangements that would lead to the same "predictions".*

[25]The fate of the GP-B satellite, once all its helium had been used up and it was effectively in a state of suspended animation (suspended in its near-Earth orbit and with limited power from its solar panels that was sufficient to receive and send signals, but no source to power its thrusters and shift its orbit significantly) was to be used as a training vessel to offer technicians of all types a real hands-on opportunity to control an orbiting satellite.

GPB: In fact, the operation was repeated 4 times: once for each of the four independent gyros, and each time a unique configuration that gave exactly the perturbation observed was arrived at.

SCI: *But it was necessary to introduce a changing polhode period into the calculations in order for the results of the TFM to agree with observation. By the admission of the GP-B team they do not understand (and it seems nobody does) how energy is dissipated from the gyro with the result that it shifts its mass distribution around its axis of rotation thereby conserving angular momentum (in an evidently unpredictable way). If they just introduce such free variables as changing polhode motion when they need to in this way they will inevitably always manage to fit the data to the "prediction" of their suitably modulated cause.*

GPB: Not only did the TFM lead to a unique distribution of frozen electromagnetic flux on the surface of the gyros that agreed with (or "postdicted" maybe) the interference in the SQUID readout signal, but that same unique solution also explained the occasional jumps observed in the spin axis orientation of the gyros as being a resonance effect that occurred when a high-level harmonic of the (varying) polhode frequency coincided with the spacecraft roll frequency; that is almost independent confirmation.

SCI: *Not at all. A method was used that was devised precisely to come up with a solution to this problem and the model it is based on has simply been expanded and adjusted until the team arrived at a result that accounted for both the observed phenomena.*

GPB: Actually, two independent techniques were used; and both indicated the same conclusion. First a geometrical approach was developed that gave early indications that there was a unique solution in terms of a joint cause with a characteristic signature that could therefore be separated out from the GTR signal. A more precise algebraic approach was then developed which required considerably increased computing power to take the analysis further. The team's resources were even divided between the two approaches and they were worked on separately to

see if they yielded different or conflicting findings; but they converged totally on a common result as far as they could proceed in tandem.

SCI: *But had the team not encountered convergence with the two methods that they invented, they would just have come up with some other method and introduced as many additional side effects as necessary that they admit they "do not fully understand"*[26] *in order to adjust the methods until they converged. Then they could have written the previous attempts off as failures and have presented these two new methods as converging and leading to a unique solution (with sufficient independent variables in their models to meet their requirements).*

GPB: In order to avoid just such an occurrence of overfitting of the models to the data at hand, methods of truth modelling were devised to ensure that the systems that were devised to analyse the actual data had not become compromised in the sense of allowing the data they were to treat to dictate how they were to treat them. Artificial simulated data with different arrangements of signals preprogrammed into them known as "truth model data" were then processed, "to ensure that evolutionary changes in the codes do not compromise their ability to correctly analyze straightforward data sets".(Kahn, 2007, p. 420) In this way the team ensured that their data treatment methods and data models were not tailored specifically to the actual data they had to process. Meanwhile, the idea that the GP-B team could have just gone on for ever without having made substantial progress is just not realistic; apart from anything else, there was the independent SAC overseeing their work and passing judgement on it.

SCI: *But it is impossible to escape the fact that the sole unique dataset was where the effects that the team went on to model were originally detected; that dataset also supplied the data they used to build the models; and now they are claiming that the same unique dataset is the evidence that they have that they have correctly modelled the interference and*

[26] This is a reference to the GP-B team's claim that energy is lost (from a "quasi-adiabatic" system!) in ways that they do not fully understand, which I quote on page 176.

have been able to separate it out to leave cleaned-up data with the GTR signal visible in it. This is clearly a case of double counting of data as both the learning set used to devise the model and the test set used to verify it: they cannot have it both ways! They have what appears to be a self-consistent theory but no evidence in its favour at all.

GPB: In fact, there are 4 separate datasets one from each gyroscope. But the important question here is: What is the chance that the data would agree with the hypotheses that the team have developed (here, effectively the TFM for each gyro) if those hypotheses were in fact incorrect? And it seems clear from all of what we have just said, not just that the answer to this question is that the chance of such an outcome would be extremely small, and therefore the fit between data and phenomenon in this case provides a very stringent test of the hypotheses; but also that the method they have adopted has tracked actual difference makers, despite not being able to go back and take more data.

SCI: *Well, maybe I'm missing something, but I am still not convinced that I should lend too much credit to the GP-B team's claims.*

In the next chapter I will see if there are further arguments that can be brought to bear to convince such a sceptical contentious individual that the reasoning of this gracious physics buff is sound.

Chapter 6

Lessons to be Learned

6.1 Introduction: The Generation of Knowledge

In this final chapter of my analysis, my aim is both global and threefold. In order to fulfil that aim I will divide the remainder of the chapter into three distinct and very separate, though I trust unified, sections. My overall aim is to situate the knowledge claims of the GP-B team within their natural and fitting context. In doing so, I hope to clarify some points and put some doubts to rest.

The threefold nature of my objective thus stems from the varied nature of the contexts within which we can situate knowledge. One of the worries that academic analysis tends to generate is that it becomes too compartmentalised and removed from the broader setting that it forms part of. So, while one of the threads of the analysis I offer here is most definitely to extend the work that I have presented in the previous chapters, the other two represent a change of emphasis and a move into slightly different, related areas.

In Section 6.2, I consider the status of the knowledge that GP-B has generated as part of an extended network of knowledge and not as isolated work or beliefs that can be judged individually in isolation from the rest of our belief or knowledge. Here, quite naturally, I rely on the foundational work of Quine and his notion of a web of knowledge. I see this as the more theoretical of the three contexts I mention.

In the central section (Section 6.3), I adopt a more practical approach. This is where I take the analysis I have offered in the previous chapters further along the same lines I have already presented. So, here I return to the central theme of the ideas of Mayo and Woodward and attempt to convince the unsympathetic reader that the position I have adopted, through the arguments I have already presented tentatively, is one that is not just an option, but compelling. As this is based to a large extent on the solutions that the team came up with in face of adversity, and hinges on the judicial application of experimental know-how and so it can be seen as pragmatic, as well as practical.

The element or aspect of the analysis that I think it is always important to include, is a social perspective. Science, and particularly physics (and even more than other areas, maybe, fundamental physics addressing foundational issues for the whole of natural science) is often seen not just as removed from the public arena, but necessarily so. I adopt an unorthodox approach in Section 6.4 and attempt to show that aspects of the debate concerning the generation of knowledge certainly should be accessible to the general public. So, though portraying the crucial judgement concerning knowledge claims as part of a court proceeding, I attempt to show how the process of knowledge generation can be more open to public debate and involvement. Thus I hope to explore three different contexts: one more theoretical, an in-depth pragmatic or practical context, and a more social setting.

6.2 The Web of Knowledge

6.2.1 Presentation

In the next section of this chapter (Section 6.3), I go on to analyse the GP-B claims from two slightly different perspectives using the requirement for stringent tests of hypotheses that was originally formulated by Mayo (SC) and which, as I explain in Chapter 4, does not necessarily coincide with a use-novelty requirement. The other vital element of that analysis is the notion that the necessary counterfactual sensitivity required between the data acquired

through experimentation and the target phenomenon, as I also explain in detail in Chapter 4, can be supplied not only by the logical relationships that predictions stand into theory, but also, at least in part, by the experimenters striving to ensure that other possible effects, causes or propensities have been accounted for or shielded against. As I have said, although it may be possible to derive certain predictions from the background theory, it is the skill and know-how of the experimenters that ensures the data actually collected can track those predictions and discriminate between the target and possible alternatives. Before I embark on that analysis, however, in this section I first want to revisit the idea I introduced at the end of Chapter 3 of regarding the GP-B results as forming part of a Quinean-like "web of knowledge" (WoK) or the "web of belief" as Van Orman Quine and Ullian, 1970, titled their book on the subject. As is well known, Quine's exposition of such a web or fabric contains several general ideas; and I think they can be employed here to see just how the GP-B team, over more than 5 decades, built up and strengthened belief in what they were doing, the methods and technologies they employed, and the new knowledge they eventually claimed to have produced. Let me just recall how Quine introduces this idea. At the start of Section VI of *Two Dogmas*, Quine states that:

> The totality of our so-called knowledge or beliefs, from the most casual matters of geography and history to the profoundest laws of atomic physics or even of pure mathematics and logic, is a man-made fabric."
>
> (Quine, 1951, p. 39)

Now, I will consider some of the qualities of that fabric, or web.

6.2.2 Characteristics of a WoK

The most important and I would say fundamental idea in the description of belief or knowledge as a web is the interconnectedness of all our beliefs and therefore knowledge. This can be seen to hold at two different levels: on the individual, personal level; and also at different collective levels, either in many different types of collaborative efforts or throughout society as a whole. It is often considered that a Quinean-type web is an exclusively internal affair

with each of us building up our own web; but for us to function effectively (without even mentioning the efficacy or efficiency of that functioning) within an extended community or maybe society as a whole, we also interconnect our beliefs with those of the people around us. In my characterisation of the WoK, I will move between these two elements, themselves clearly intertwined, as most of what I say applies to both the internal and social levels, although some it can be seen to suit better one or the other.

When Quine first published the notion,[1] it was in the context of his criticism of reductionism within scientific thought and knowledge.[2]

He wanted to move away from the idea that individual statements are—or even *could* be—deemed true or false individually through comparison with experience. His aim was to move towards the holistic view that knowledge

[1] In a footnote to the title of the paper, Quine tells his reader that: "Much of this paper is devoted to a critique of analyticity which I have been urging orally and in correspondence for years past." There is no parallel mention of the second of Quine's dogmas: reductionism. Seeing as it is in the part of the paper dedicated to his critique of this second dogma that he talks of the web or "fabric", I assume that this was not just the first time that he published his ideas on this subject, but the first time they became widely known (as opposed to his already well-known view, within his close philosophical circles at least, of analyticity).

[2] After centring his criticism on analyticity throughout the first four sections of the essay, in Section V of *Two Dogmas*, Quine switches the focus of his criticism to the notion of reductionism, which he initially characterises in a (long-gone, it would seem, even back in 1951) extremely "radical" form. Then, within the context of the last two sections of the paper dedicated to criticising epistemic reductionism in general, he tells us:

> The dogma of reductionism survives in the supposition that each statement, taken in isolation from its fellows, can admit of confirmation or infirmation at all. My countersuggestion, issuing essentially from Carnap's doctrine of the physical world in the *Aufbau*, is that our statements about the external world face the tribunal of sense experience not individually but only as a corporate body.
>
> (Quine, 1951, p. 38)

forms an (ideally self-consistent[3]) network of interconnected beliefs. So, according to this view, when we experience something new, acquire additional knowledge and especially change our opinion regarding some or other proposition, we are required to adjust the entire fabric of our knowledge system to allow for the novel disposition of the whole. Famously, in what appears to me to be one of the phrases that is quoted most often from the philosophy literature of the 20th century, he claimed that our belief system or total knowledge "impinges on experience only along the edges" (Quine, 1951, p. 39). The simile he then goes on to use, which I find even more enlightening, is not so famous: "Or, to change the figure, total science is like a field of force whose boundary conditions are experience." He continues:

> A conflict with experience at the periphery occasions readjustments in the interior of the field. Truth values have to be redistributed over some of our statements. Re-evaluation of some statements entails re-evaluation of others, because of their logical interconnections—the logical laws being in turn simply certain further statements of the system, certain further elements of the field.
>
> (Quine, 1951, p. 39)

So, here we have the first characteristic of the Quinean-like WoK that I wish to consider: knowledge is holistic in nature, totally interconnected and any revision affects (what we may at times consider to be[4]) other elements of the web.

The second characteristic that I wish to consider is to do with the relative situation or position of different elements within the interconnected WoK.

[3] Although Quine (and Ullian) insists on the requirement of consistency and uses this as a primary motivation for revising our beliefs, it has since been shown that in fact our systems of knowledge typically admit inconsistencies without our feeling the need to revise them. In connection with a social WoK, societies always have inconsistencies in their overall beliefs, but it is the nature of social interaction that we tolerate certain inconsistencies and shun others.

[4] There is an inherent tension throughout this analysis insofar as if we wish to consider the web in a holistic fashion, the division into "elements" can only be a methodological device to facilitate our study. I will not go into this issue here; suffice it to say that when I talk of elements, I am not suggesting that we can actually separate one part of the web from the rest of it: that would be contradictory to the notion of the web, at least in terms of the truth value or status afforded each such element.

Returning to Quine's own words again, we find that he tells us that we have a "natural tendency to disturb the total system as little as possible". Consequently, we are more likely to revise statements that:

> . . . are felt, therefore, to have a sharper empirical reference than highly theoretical statements of physics or logic or ontology. The latter statements may be thought of as relatively centrally located within the total network . . .
>
> (Quine, 1951, p. 41)

than the more "peripheral" statements that are almost directly impinged upon by experience. Thus we can see an overall structure emerge. The fundamental pillars of all our scientific knowledge, famously logic and arithmetic,[5]

stand at the centre of the web; with many of the basic tenets of our current physics theories not that far removed from them.[6] The most readily revisable statements that rely directly on the sense-data that we perceive are on the very edge of the web; including every direct observation of the world around us and each individual measurement taken in a laboratory. The rest of our knowledge is arranged along a radial gradient of our readiness to revise it: the more a statement depends on direct empirical evidence and therefore

[5] Although these are seen as forming the central structure of the web, it is important that the third characteristic of all the (elements of the) web that I consider below (see the quotation on page 193) also applies to them; that is, that they too are in theory revisable. As an example of explaining this, in his recent study of the place of logic within the structure of Quine's WoK, Matthew Carlson tells us:

> It is consistent to hold that, while we cannot currently make sense of revising logic, we may nevertheless someday be able to do so. Quine's claim that logic is in-principle revisable just amounts to a refusal to rule out, on philosophical grounds, any future course that may be taken by 'scientific method, unsupported by ulterior controls.' It amounts, in other words, to a refusal to predict—or prescribe—the future of science.
>
> (Carlson, 2015, p. 7)

[6] Again, we can see this at the level both of individuals and of society as a whole or communities within broader society. Whereas individuals certainly do hold very firm beliefs regarding logical conceptions and constructions (or views regarding physicalism for example), the standing of entire theories is much more a matter for collectives or society.

the less us revising it would disturb the rest of the web, the further out it is located. In contrast, the more well-entrenched a belief is within our entire system of interconnected beliefs, and the more it forms part of or is required for other beliefs, the closer it is to the centre of the web. Furthermore, due to the interconnectedness of the web, these relative positions are reflected in the potential connections leading to and radiating from each element in the web. Thus, the more connections any specific belief or piece of knowledge has to other pieces of the web, the more central it is; and the more costly it would be to revise, as such revision would cause greater upheaval throughout the rest of the web.

The third and final characteristic of the WoK that I want to consider (maybe the best known and most often cited by philosophers) is summed up succinctly by Quine thus:

> Any statement can be held true come what may . . . by the same token, no statement is immune to revision.
>
> (Quine, 1951, p. 40)

In this way Quine wishes to make it clear that the holistic nature of the WoK means that we always have options as to what to revise and adjust. Even central elements could be revised (as I note in footnote 5 above), but as one of the previous quotations makes clear, our preference is always to adjust the edges and leave the central pillars of knowledge intact, due to the immense readjustment in the whole of our belief system that would be required to declare a central element false. To illustrate this, in his encyclopaedia entry on Quine, Peter Hylton offers us the following:

> The truths of elementary arithmetic are an example: they play a role in almost every branch of systematic knowledge. For this reason, we cannot imagine abandoning elementary arithmetic. Doing so would mean abandoning our whole system of knowledge, and replacing it with an alternative which we have not even begun to envisage. Nothing in principle rules out the possibility that the course of experience will be such that our present system of knowledge becomes wholly useless, and that in constructing a new one we find that arithmetic is of no use. But this is a purely abstract possibility, certainly not something we can imagine in

> any detail. So the idea that we might reject arithmetic is likewise unimaginable.
>
> (Hylton, 2014)

Armed with these three characteristics of a Quinean-like WoK (holistic interconnectedness; gradient of reversibility from most revisable at the edge, to least at the centre; and no fixed, irrefutable knowledge, even at the centre), I can continue below to consider GP-B and its relevance for GTR from such a perspective.

6.2.3 Levels within or perspectives on the Quinean-like WoK

Before embarking on my analysis of GP-B using the perspective of the WoK structure, I just want briefly to consider what I will term different sub-webs of such a web. In Two Dogmas, as I have said, Quine originally talked of a fabric of beliefs; this was then developed further in *The Web of Belief*. Here, instead of just considering individual beliefs or propositions, I am now interested in what we could consider to be a different layer of structure (maybe sub-structure) of sub-webs that exist within such a WoK: that of whole theories that different clusters of beliefs could be seen as forming. A cluster may of course amount to more than the sum of its parts, and I want to consider that complete scientific theories (and also our technological know-how) can be seen as introducing a new layer or level of sub-webs to the WoK; despite relying on and being built from all the individual beliefs that go into the theory (or our technological know-how), a theory can itself be seen to stand in certain relations to other theories as an element of a network consisting of different sub-webs within its overall structure.

Just as individual beliefs, typified by what we consider to be true propositions, and the knowledge that we derive from them are all interrelated and mutually connected, so too are our scientific theories that we derive from the conjunction of many different beliefs. This can be seen as a type of superstructure superimposed on the WoK, gathering together many different strands to form a network of theories within the web. Just as with beliefs concerning individual propositions, we can see the same web or network

structure reflected in our far more complicated theories; they share the same crucial three characteristics I describe in the previous subsection, being also connected to the individual beliefs from which they are formed. So, the closer one theory is to the centre of this superimposed network, the more it grounds or is required for other theories. In this way, well-established fundamental theories are meshed together within an overarching central structure of the WoK. They in turn form a base for the more recent additions that stretch further out in different directions, representing the different fields of our knowledge (that seem so isolated at times, as I mention in the Introduction to Chapter 2). As we continue to move out, we come to still newer theories, which may be connected via many links to other theories, but which are now becoming more tenuous and more readily revisable (that is, it would be less costly to the structure of the totality of our theories to revise them) as we are further removed from the solid central structure of theories (at this superimposed structural level). Finally, as we approach the edge of our WoK, we encounter the peripheral, uncertain, newest areas of knowledge and representations of nature. Here are the theories that are still trying to establish firm ground on which to stand via connections to all the other knowledge that is around them. Once, again, this can be seen to work both at the personal, individual level and also at the level of collectives and society as a whole.

I introduce this notion of the WoK housing and giving rise to different structures of sub-webs here, simply to make it clear that what holds in terms of the characteristics of individual beliefs, typically in the form of single propositions, can also be seen to hold for entire theories. In what follows, when considering GP-B, I will at times be interested in these different dimensions of knowledge, and at times the connections both within a sub-web (theory, consisting of limitless individual beliefs) and between the different sub-webs (the dependence of one theory on another). At one level, GP-B was envisaged and designed to test GTR: to place or refine limits on our confidence in the entire theory. On a less grandiose scale, due to the problems encountered during the mission and data analysis, the results and the claims that the GP-B team reported rely on the status of their method of trapped flux mapping (TFM) as a completely new technique for analysing electromagnetic flux on spinning

superconductors. We should also consider the status of our knowledge concerning the new technologies that were developed specifically for the GP-B project or which were adapted and co-opted into it. As I mention in Chapter 3 (footnote 17), the team claimed to have developed nine new technologies which were incorporated into the hardware of the project. Meanwhile, maybe more at the level of individual propositions, we have each individual reading taken on board the actual GP-B satellite and the status of our confidence in those data. It is necessary to consider these different levels, or aspects of our knowledge: individual propositions, and whole techniques, methods or theories.

6.2.4 GP-B seen from the perspective of a Quinean-like WoK

Now the question is: where can we see (or should we place) the knowledge provided by GP-B within our WoK? As I have just said, I consider it possible to describe knowledge as being on different levels within the WoK, or pertaining to different aspects of it. So, I will consider how we can describe whole theories, technologies or certain techniques each as a different sub-web which in turn forms part of a network of sub-webs both within and across the different levels or aspects of the global WoK. In this way, I analyse how the knowledge generated by GP-B may be seen to fit into our WoK in different ways related to different levels of knowledge or aspects of the WoK.

First I consider the level of complete theories. Aspects of our fundamental physics theories—quantum physics and GTR—are well established within the realms to which they apply. However, we are well aware of the incomplete and transient nature of these theories. Indeed, the project of reconciling their differences and providing a single theory of quantum gravity seems, at the moment, to require innovation on a scale that thwarts our best attempts. Nonetheless, at least parts of those theories relating to the effective modelling of relations that hold between elements of the external world (though certainly not the interpretation of the theoretical constructs used to try to explain the efficacy of some of those models) are a firm base on which to build. It is precisely the fact that they both provide us, time and time again, with such

robust results that makes it so difficult to see how they may need to be adjusted in order to iron out the contradictions that exist between them. Thus, we see GTR as a solid central part of our belief system or WoK. Much current technology—such as, maybe most famously, the GPS system used for navigational purposes—and the interpretation of virtually every astronomical observation made today (beyond those of near-Earth objects) rely on our understanding of gravitation as described by GTR. Its status as a central part of our current physics and our whole WoK, which to some extent is constantly being confirmed by the use we make of it in everyday technological applications, thus seems to be extremely well founded.

The aim of GP-B was of course to bolster, or question, this status through very specific scientific experimentation. Had the results of GP-B not agreed with the predictions of GTR it would have resulted in considerable epistemic turmoil. The claim of the GP-B team is that it did indeed confirm GTR (and thereby also the adequateness of a metric representation of gravitation; and of course, EEP) to a more accurate degree than any other experimental findings at the time via its 18% error margin on the frame dragging measurement and (limiting all observations of massive bodies, as opposed to measurements involving only massless phenomena) 0.3% on the geodetic prediction. The effect that this has on our belief in GTR cannot be great: we were already convinced (as a society, if not as individuals!) of its validity, at least to the degree that GP-B tested it. Of course, to the extent that it represents the accumulation of evidence in favour of GTR, or the refining of the limits of it, the results must increase our confidence in GTR; but we are dealing with a central and remarkably solid element of our WoK. GTR is revisable, as all our knowledge must be, and the agreement of the results of GP-B with the theory must therefore make it potentially more costly to revise its status. We must therefore see the theory post-GP-B as being more firmly anchored to the central part of the WoK (even if the increment is small).

Be that as it may, I have not yet considered the global interconnectedness of the different aspects or levels of our WoK. Despite being a (straightforward, in some senses) test of GTR, with the relatively central position within our knowledge structure that that theory holds, the knowledge we gained from GP-B

rests on new technologies (and analytical methods) with few connections—as yet—to other knowledge. This can be seen from the perspective of a different level or aspect of the WoK. Just as in the case of the dimension of whole theories, we can see each technology or area of technological know-how as being an element or sub-web at the corresponding level of the WoK. The GP-B result has myriad tentative connections to technological know-how which is itself still situated in outlying regions of our WoK. It remains to be seen whether these connections will strengthen over time and with application, and the GP-B results will thereby become a more firm part of our knowledge structure; or whether, alternatively, one or more of the connections will prove to be a weak link that shakes the results of GP-B, and just what consequences that would have. So, in order to anchor the top level or aspect of entire theories of our WoK, which could be seen as GP-B providing stronger grounds for confidence in GTR, we need to consider the dimension of the technologies involved at arriving at the results; and their status as elements in the corresponding level of knowledge. As an example at this level, I now consider the fused quartz bonding system that was developed for GP-B to form the necessary bond between the SIA (scientific instrument assembly) and the on-board telescope.

As a new technology developed for the project, we must see the system developed by the GP-B team to bond fused quartz and our knowledge concerning it as forming a new limb of our WoK: before it was developed by the GP-B team we knew nothing about it. Through laboratory testing, of course, we initially learned about it and established it among the technologies we have at our disposal. It must still occupy a peripheral place in our WoK though. It is only though continued and varied experimentation and application that we can strengthen its ties to other aspects of our knowledge, and move it gradually nearer the centre of our WoK if our beliefs concerning it become more connected. In the case of the fused quartz bonding system developed for GP-B, such a shift towards the centre has indeed been achieved over the years since it was first developed. We are told by the GP-B website that nowadays: “Industry applications include bonding improvements in optoelectronics, precision optics, laser optics, laser crystal augmentations, general

optomechanical applications and creation of optical systems."[7] This pattern can be seen repeated in many of the new technological applications that were developed for GP-B. Maybe most famously, the refining of the GPS system to provide the degree of accuracy required for GP-B has since been applied to automated tractor control and to landing aircraft automatically. Thus, each of these elements could initially have been seen as forming a (new) outer limb of our WoK; but each has been shifted towards the centre through applications in different fields and continued use and development, and the concomitant increase in our knowledge concerning each technology.

Due to the interconnected nature of all our knowledge, this must ultimately be reflected in the status of the GP-B result within our WoK. That result relies on all the different technologies that were employed throughout the mission, many of them novel—such as the entire drag-free satellite control system which has since been used time and time again in satellite applications—and every time they are successfully employed in other fields, they become more connected and shift further towards the centre of our WoK. This of course has a transversal effect: not only does the technological application (a sub-web, within the technology aspect of the WoK, as I have referred to such elements) move inwards, but so does the knowledge at different levels of the WoK that is connected to that technology (crucially here the GP-B findings and therefore GTR). However, that cannot be seen to be the case with the TFM developed by the team. TFM is neither a technology that has broad and varied applications, nor a well-grounded scientific theory in the way that we can consider GTR.

Through this new data processing technique that the team were forced to develop to overcome the severe restrictions imposed on them by the nature of the actual data they collected, the GP-B result has certainly been drawn further from the central structure of our WoK. Indeed, it may appear that whatever other connection there may be, all the team's claims hang by the thread that represents these new, otherwise untested, data processing techniques and theories.

[7] See https://einstein.stanford.edu/content/spinoffs/spinoffs.html accessed on 27-07-2016.

Of course, this is not necessarily a bad thing—for human knowledge; it certainly was for GP-B!—and is to a degree inevitably how we build up new knowledge. Slowly, as we become more confident in the work on which experimental results rest (through increased use and agreement with prediction, or adjustment to fit other new knowledge) so they become more firmly established as part of the structure against which to test new hypotheses. Inevitably, experiments such as GP-B, which may be conceived as truly historic, groundbreaking steps, must take their place within our broader framework of knowledge. The results do not stand alone, but rely on an elaborate interdependent system of observation, interpretation, modelling and theory.

From a distance, with a historical perspective, it may well be the case that we mistakenly attribute too much importance to an individual experiment. On further investigation and reflection, such crucial experiments invariably turn out to form just one part of a larger picture which includes many factors leading to an important change. This, for example, is often considered to be the case of the Michelson-Morley experiment. It is cited as the groundbreaking experiment that marks the end of the aether model and the beginning of the road to relativity. On further study of the case, it can be argued that, at the time, it was far from the decisive step so many have claimed it to be. In fact, it was considered by Michelson to have been a failure and is not believed to have been influential at all in the genesis of STR, since Einstein does not seem to have been aware of the results until after 1905. At the time, it was very uncertain and on the outskirts of our WoK. It was only as connections were made with other knowledge that it formed the connections necessary to be considered a historic breakthrough.

That may well be the case with GP-B; we simply do not know at the moment. What type of reception the GP-B results will receive from future generations we just cannot tell. They may gradually receive increased confidence as the work on which they rest is revalidated. The results may be revisited and reinterpreted, and may even contain information that we have so far been unable to interpret. Alternatively, the results may be thoroughly revised in the coming years due to our changing knowledge base and the confidence we deposit in new and different findings. This is the way to build a solid

knowledge base and whatever happens, we should make sure that we are prepared and willing to constantly re-examine previous experimental results, as with all elements of our WoK, and also to have the confidence in them that they deserve. Having now considered this more theoretical perspective on the process of the generation of knowledge through GP-B, I will now move on to more practical or pragmatic considerations of that process.

6.3 Counterfactual Double Counting

6.3.1 A practical approach

As I state in Chapter 4, it is often claimed in the philosophy of science that some form of "predictive novelty" on the part of hypotheses adds weight to the support they gain from their agreement with data. Here, for example, is one succinct statement of such a view:

> When a scientist uses an observation to formulate a theory, it is no surprise that the resulting theory accurately captures that observation. However, when the theory makes a novel prediction–when it predicts an observation that was not used in its formulation–this seems to provide more substantial confirmation of the theory.
>
> (Hitchcock and Sober, 2004, p. 1)

Coming as it does from seasoned philosophers, this is a very broad statement that avoids all talk of necessity. However, as I show in Chapter 4, a stronger view which is also common is that such use novelty (UN) is a necessary condition for experimental data to count as evidence in favour of a hypothesis they agree with (or match in some appropriate way). In other words, on this stronger view: if experimental data are used in the formation of a hypothesis, then the agreement of that hypothesis with those data can in no way count as evidence of the truth of the hypothesis in question. The formulation I give in Chapter 4 (page 130 is:

> the evidence must not have been used in the construction of the hypothesis.

Source	$\mathbf{R_{NS}}$ **(marc-s yr^{-1})**	$\mathbf{R_{WE}}$ **(marc-s yr^{-1})**
Gyroscope 1	-6588.6 ± 31.7	-41.3 ± 24.6
Gyroscope 2	-6707.0 ± 64.1	-16.1 ± 29.7
Gyroscope 3	-6610.5 ± 43.2	-25.0 ± 12.1
Gyroscope 4	-6588.7 ± 33.2	-49.3 ± 11.4
Joint	-6601.8 ± 18.3	-37.2 ± 24.6
GR prediction	-6606.1	-39.2

TABLE 6.1: Final GP-B results.
[SOURCE: Everitt et al., 2015]

Here, in GP-B, I claim that we have a case where this UN condition can be said not to have been met. It can most certainly be argued that one and the same unique dataset was used both to develop the novel theoretical TFM and then to show how that theory together with the dataset are evidence in favour of GTR. Yet the GP-B team do not appear to see this as problematic: they claim to have used their TFM as a tool in their confirmation of GTR. Of course they accept that this extra level of modelling and calculation increased the uncertainties in the final results, but there is no questioning of the validity of those—less accurate than desired—results. Let me just reproduce those results as they appear in the detailed introductory paper to the 2015 Classical and Quantum Gravity Focus Issue (Table 6.1) and what Everitt at al. have to say about them:

> Table [6.1] gives the result for each gyroscope in marc-s yr^{-1}. Within the one-sigma limit they agree and confirm the Schiff geodetic and frame-dragging predictions to 0.3% and 18%.
>
> (Everitt et al., 2015, p. 23)

If, as I do, we accept that this is a case of double counting (that is: non-novel use of data), then the GP-B team has implicitly rejected a UN requirement (although, of course they may (silently) accept the observation made by Hitchcock and Sober, 2004, quoted above, and simply see their findings as not providing the coveted "more substantial confirmation"). The issue that I am interested in here is whether we are facing an over-optimistic claim of having arrived at results that actually support GTR, when in fact, there is no severity

to the test that the match between data and GTR represents, because the two could not have failed to match. This latter possibility would be the judgement brought to bear by someone who sees UN as a requirement, and not just as a way to provide "more substantial confirmation". For that critic of GP-B, the fact that the dataset were used to arrive at the hypothesis (in this case: the "cleaned-up" GP-B results agree with the predictions of GTR) means that the claim that the results agree with GTR simply represents one—incredibly complicated—possible interpretation of the situation. According to this interpretation, although the team have come up with a self-consistent explanation of the results (it may well be the case that, in fact, the TFM method and findings are correct, this critic would concede), there is no independent evidence that this is indeed the case.

Drawing on much of the work I have already laid out in Chapter 4, I will argue that this episode is one where, by adopting Deborah Mayo's (1991) severity criterion (SC) for the validity of the tests that theoretical hypotheses pass when experimental data agree with them, the experimental results eventually reached can be seen to be perfectly justified. This is despite the scepticism that NASA alluded to back in 2008 (see my discussion of the NASA, 2008 report in Section 5.1). This has to be shown though the specifics of the experimental set-up and execution: having "failed" a UN test (if there were such a thing), the onus is on demonstrating that the route adopted to arrive at the results has adequately ruled out the vast majority of other possibilities. In other words, despite the data standing in the relationship of entailment that all episodes of double counting necessarily involve, here we have an occasion where the specifics of the experiment mean that *nonetheless* the data are still tracking the genuine difference makers. Thus, the scepticism is not justified.

In this way, I consider this to be a more practical approach to analysing the claims made by the GP-B team when they published their final results confirming GTR. I say that it is practical because it is an innovative and pragmatic response through the actual practice of experimentation; it demonstrates how new knowledge claims can be justified, and in this case were, despite possibly failing to meet some more theoretical criteria (such as UN).

6.3.2 Mayo revisited

Episodes such as this (where the impossibility of refining, improving or repeating an experiment leads us to consider novel ways of extracting information from a unique dataset) demonstrate the importance of Mayo's SC in the practice of scientific experimentation and the concomitant generation of knowledge. As Mayo says in her 2008 paper where she defends her SC from criticism, her:

> goal is to set the stage for philosophical scrutiny into the cases still under dispute in practice.
>
> (Mayo, 2008, p. 872)

GP-B is just such a case and what I present here is some scrutiny of it in the light of Mayo's SC. Let me just recap the main points of Mayo's SC and what is required for the match between data and prediction to be a severe test of the hypothesis under scrutiny in cases where a UN criterion is not met.

First, to be an instance of double counting, we have that data, x, stand in the relation to the hypothesis constructed using those data, H(x) (and which the data are said to be evidence in favour of), as follows::

> The bare bones of a use-constructed test procedure is to output H(x) as supported, well tested, indicated, or the like by data x, where x has been used to construct . . . H(x) in such a way as to fit, or pass, or be in accordance with, x.
>
> (Mayo, 2008, p. 859)

This quite definitely seems to conform to the situation with respect to the TFM: the GP-B dataset (x) was used to construct the TFM for the gyroscope (H(x)) which is claimed to be well tested by x.

The severity is then given by:

> with very high probability, test T would have produced a worse fit with H (or no fit at all), if H were false or incorrect.
>
> (Mayo, 2008, p. 860)

Here, apart from the match between x and H (which is guaranteed in cases of double counting) we have T, which is the test that the specific experiment represents. This is the step that can be missed if data and phenomenon are

conflated, and only the logical relationship between the background theory and the predicted phenomenon is considered; again, as I explain in detail in Chapter 4. The crucial point here is that if H (in the case of GP-B, we can consider this to be the TFM) was false, it would be extremely unlikely to have been able to arrive at the match between H and x; and this is determined by the specifics of the the test, T, that the experiment represents. This means that:

> When the severity requirement is satisfied, it is because the falsity of H would render the fit (between H and x) extraordinary
> (Mayo, 2008, p. 861)

So, I hope to have made Mayo's SC and its difference from a UN criterion clear; together with the fact that it is the specifics of the experiment that make the crucial link between data and phenomenon. I can now consider how the GP-B data can be seen to be combined with the appropriate counterfactual sensitivity provided by the specific experimental set-up and execution, to produce severe tests of the underlying phenomena.

6.3.3 GP-B data production

As I explain in Chapter 5 above, the first important breakthrough in being able to treat the "noisy" or "dirty" datasets resulting from unexpected perturbations in gyro motion and salvage from them the relativity signal that was the target of the mission, came with the identification of the underlying cause as electromagnetic in origin. There was only one unique science dataset to work with. However, there was also the calibration data which proved so useful in identifying the cause as electromagnetic. Since the calibration data gathered after the Science Phase of the mission had concluded was separate from the main dataset containing the science data, I will not consider it here: I accept that it demonstrated the electromagnetic nature of the perturbations. There were also the extremely detailed engineering "snapshots" of data which provided 2-second bursts of high-resolution information on gyro orientation. So, as I have said, if a cause could be found and its trace had some distinguishing feature, then there was a chance of being able to identify the contribution to the total signal due to this unwanted cause and subtract it out to leave a

"clean" dataset. It was vital that the interference was not mirroring the GTR effects (gradually pushing the gyro out of line over the experimental period) or behaving in a way that was similar to the known perturbations: acting at the spacecraft roll rate, for example. This is a crucial claim of the data processing results.

One of the key concepts in the argumentation here is that of data, in terms of exactly what is produced, gathered and processed. Data are what I will employ as evidence of certain underlying phenomena, which are held to be responsible for certain features of those data. So it is vital that we are absolutely clear as to what is being produced, gathered or measured, and then processed. Throughout this section—and indeed in the vast majority of this chapter; and the entire thesis insofar as it deals with GP-B—what I am interested in and what I am talking about is the SQUID output signal. That signal was a voltage. All the information that was used by the GB-P team to develop their TFM method came from that voltage readout. Of course, it had to be combined with data from many other sources, but essentially all the information concerning how the gyros were responding came from the SQUID readout data. So, the confirmation of the predictions of GTR (to within approximately 18% in the case of frame dragging and 0.3% in the case of the geodetic effect) also came from the SQUID output voltage. It is in this sense that we can see this as an example of double counting: the SQUID output voltage signal was used to develop the entire TFM method, draw up the precise maps of the specific flux trapped on each of the gyro surfaces and then to predict the part of the voltage signal that was due to the trapped flux (which was then subtracted out from the signal to arrive at the final result). Below I will consider the counterfactual reasoning that takes us from the output signal to the confirmation of GTR, but first let me recap the details of the SQUID output signal, the data, in more detail.

The SQUID output voltage signal was produced primarily by the LM of the spinning superconducting gyro shells. However, as I explain in Chapter 5, there was also the unexpected contribution from the trapped flux, as it rotated and interacted with the uneven distribution of electrical potential surrounding it resulting in the interference from the so-called patch effects. The crucial

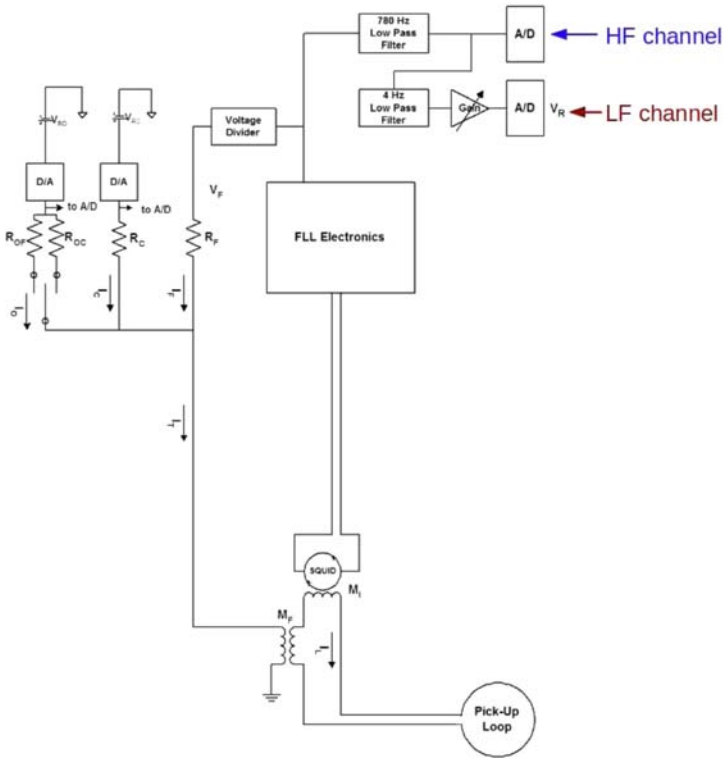

FIGURE 6.1: This is a very basic representation of some of the components in the SQUID circuitry that intervene to pre-treat the signal before it is measured.
[SOURCE: Silbergleit et al., 2015b]

issue concerning these data is just how and what the on-board instrumentation was actually measuring. If we have a steady signal, a voltmeter will provide us, to within the accuracy of the apparatus, a constant reading; if we have a varying signal, as in the GP-B case, the reading will vary and by recording the signal what we are effectively doing is sampling the signal at different time points. In the GP-B satellite, this sampling was performed and the corresponding measurement was recorded at 2 different frequencies. As I explained previously, the target science data which GP-B was designed to produce and was recorded throughout the whole of the Science Phase of the mission, was a sinusoidal voltage wave whose period was the roll period of the spacecraft, 77.5 seconds. To produce the final GTR result, all the information collected in each GSV period (approximately 60 minutes of every 98-minute orbit) would be condensed into just one data point, which

reflected the direction in which the gyro was actually pointing during that GSV period. However, the information contained within the sinusoidal wave itself was also vital. As I say in Chapter 5, it was used to calibrate the on-board equipment. So, in order to be able to reproduce the sinusoidal wave accurately, these science datapoints were collected, that is the SQUID output voltage was sampled, twice per second (the LF signal I refer to in Chapter 5 with an effective frequency of 0.5 Hz)[8].

When we talk of and report the GP-B data, we usually do so in terms of angles, not voltages. As I hope is clear from previous chapters (Chapter 3 and Chapter 5), this conversion should have been a relatively straightforward process. Controlled "dither" was introduced into the spacecraft motion to calibrate the SQUID output signals as the craft moved through known angles. Furthermore, effects of both orbital and annual aberration on the difference between the telescope pointing direction and the actual inertial direction to the guide star can be calculated extremely accurately and its effects are clearly visible in the science data signal for each orbit and over the year of the experiment. This allows further calibration of the equipment to convert the voltage output into an angle. This task was also complicated by the interferences in the actual signal, but the theory behind the conversion remains essentially the same and we can effectively talk with equal validity of either the actual SQUID output voltage or the pointing angle that the output voltage represented.

As I say in the previous chapter, the main problems with the LF signal were twofold: it was extremely noisy; and there were "jumps" in the signals, of the order of 10-100 mas, over several days (see Figure 5.2 in Chapter 5). However, the 2-second long snapshots containing much more detail were used to develop the TFM. These data were also the SQUID output voltage reading, but in this case, instead of sampling the signal just twice per second

[8]Of course this is itself an idealisation, on several levels. As I explain in Subsection 5.3.2, the signal was actually recorded at a higher frequency, outliers removed and the different readings combined to give a preprocessed average for this value. Again, the idea that a reading is an "instantaneous" value of the tiny current induced in the pick-up loop is another idealisation. Moreover, the SQUID output reading has already undergone considerable pre-treatment via different electronic components, as can be seen from the basic circuit diagram included here as Figure 6.1

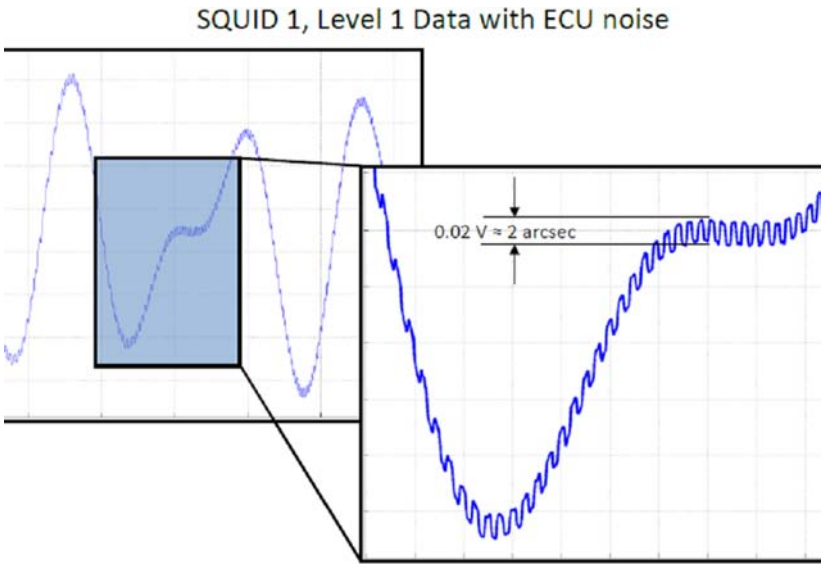

FIGURE 6.2: This figure shows how the SQUID output voltages were effectively converted to angles and the two types of measurement commonly combined. Once the conversion factor was determined (and it was found to be constantly changing, so the process was far from simple) the (short-lived) relation meant that talk and depiction of voltage and angle could be readily combined. (Here the source of the interference was identified as an experiment control unit (ECU).)
[SOURCE: Silbergleit et al., 2015b]

(in order to plot the sinusoidal wave with a period of 77.5 seconds, therefore each sine wave consisting of 155 datapoints), the voltage was recorded some 2,200 times per second: with a frequency of 2.2 kHz. These incredibly detailed engineering data consisted of many datapoints for a single rotation of each gyro (which were spinning at frequencies of around 80 Hz). So the 30 odd data points per rotation, continuously over some 160 rotations (in each 2-second snapshot) is what allowed the GP-B team to observe the variations from the target smooth sinusoidal science data in detail and develop both the TFM and establish the rate at which a changing polhode period was modulating the data. (See Section 5.3 for further details.)

The calculation of the TFM and the changing polhode frequency is extremely complex and I have gone into it in as much detail as I will in Chapter 5. Here

I want to make the reasoning adopted by the team explicit and consider the chain of warrant leading from their actual observations to their final claim and whether it can be considered a stringent or severe test. I wish to relate the procedure that was undertaken to the severity of the tests that a hypothesis may be subjected to and thereby consider whether the claim that "the results are in agreement with general relativity" (Adler, 2015, p. 1) is wildly optimistic, considering the interference encountered and the procedures adopted, or is a reasonable conclusion given the chain of argument and reasoning. Finding themselves with a noisy, one-off dataset, the GP-B team identified a potential root cause and traces of two specific effects within the data. They worked backwards from the distinctive signature of the interference recognised as (relatively) large classical torques acting on each gyro which were electromagnetic in nature and irregular polhode motion (resulting in energy dissipation), and linked both to electrostatic patch effects. As I explain in the previous chapter, using the perturbation, they modelled the magnetic flux trapped on the superconducting surfaces of the gyroscopes believed to be the root cause and calculated the resultant interferences for the entire mission. Essentially their claim is that this allowed them to arrive at valid confirmation of GTR.

6.3.4 Two problems, one cause, one claim

The reasoning adopted here can easily be divided into two different parts: one related to the team's claims regarding their cleaning up of the noisy dataset and the other to the original target of the experiment. The overall claim of the GP-B team is thus be seen as twofold: that they successfully subtracted out the unwanted signal from the science data to leave them with "clean" data and that the precession of the gyroscopes that can be gleaned from that clean data is a consequence of effects predicted by GTR. The first part is the claim related to which NASA cited scepticism within the scientific community (NASA, 2008, p. 25). That was a criticism that was levelled at the project in the early days of the data analysis when the scale of the problems had been identified but the proposed solution had not yet been worked out. One thing that was well known about GP-B was that it was an unrepeatable experiment and that the unique dataset obtained was unacceptably noisy. It seems perfectly

reasonable to be sceptical about the prospects of coming up with useful results given that situation (indeed, it seems to be positively desirable that a healthy dose of scepticism should be applied in such circumstances, in keeping with Merton's canonical norms for scientific practice (Merton, (1973)[1942])). If we also allow for the widespread dismissal of the idea that any non-UN use of data can count as a stringent test of a hypothesis, it seems natural (though not necessarily justified) that this unique dirty dataset was doomed to be written off as an extremely expensive total disappointment.

Before looking at that vital claim in more detail, let me just comment on the second part of the overall claim, concerning the original target of the experiment. There are three major theoretical suppositions that lead to the possibility of imagining a gyroscope experiment to test GTR, as Schiff and Pugh did back in 1959-60. One, which I discuss in Chapter 3, is that the spin axis of a gyroscope in free fall is parallel transported around a geodetic path. Another is that such free fall is universal; that is, that EEP (discussed in Chapter 2) holds and gravitation can be described by a metric theory. Finally, the piece that brings these two together and makes GP-B a test of GTR and not just of the metric nature of gravitation is that the specific details of the motion of a gyroscope around a massive spinning body are described by GTR within the specific metric first reported by Lens and Thirring, incorporating both the previously reported de Sitter geodetic effect and also frame dragging. When reviewing the theoretical basis of the experiment, we are assured by Adler that:

> each of these three elements ...is solidly based on previous experiments and well-tested theory. The agreement of GP-B with theory strengthens our belief that all three elements are correct and increases our confidence in applying GR to astrophysical phenomena.
>
> (Adler, 2015, p. 1)

This can be expressed as the counterfactual reasoning that actual experimenters typically use to track difference makers and confirm (or not) our beliefs concerning our representations of underlying phenomena; which I now do.

- Had the data collected during the GP-B mission not revealed a combined precession equivalent to that predicted by GTR to originate in the so-called geodetic effect and frame dragging, then we would not be in a position to claim to have confirmed, via a combination of free fall parallel transport around geodesics and the fact that gravitation is a metric phenomenon, that GTR provides an accurate description of this phenomenon.

This may seem like a very straightforward statement of experimental results, and indeed it is intended to be. Its purpose is to make explicit the connection, via counterfactual reasoning, between the data collected and how they identify the difference the underlying phenomenon in which we are interested makes to them; and also how the phenomenon is predicted from the overarching theory we use to represent it. As Adler again tells us, if this had not been the case, "a major theoretical quandary would have occurred." (Adler, 2015, p. 1)

However, I wish to concentrate on the claims made regarding the perturbing signal and the procedure adopted to clean up the data. First the dataset was used as evidence for an unexpected, perturbing phenomenon that introduced so much interference into the data that the uncertainty in the results makes them useless for refining the target parameters beyond previous levels. Then, the same dataset, that is the SQUID readout, was used to create a model of the effects of the postulated perturbing phenomenon: through TFM. Finally, the model created from the dataset in this way was used to clean up the dataset and leave the target signal visible within it. One of the crucial points in interpreting the results of this process is that the same process was applied to each of the 4 gyroscopes separately and in each case it yielded satisfactory results. Had the model been over-fitted to the available data (that is, if the model that the team devised had been merely a description of the actual dataset and not tracking the difference the underlying phenomenon was making), then we could not have expected it to produce useful results from the 4 four independent configurations: one for each gyro. I will now list what I consider to be the main claims involved in this chain.

- The short (2-second) bursts of high-frequency (collected at 2.2 kHz) engineering data were sufficiently detailed and long enough to distinguish

within them the periodicity and magnitude of perturbations; and also that this same HF signal contained sufficient data to track these perturbations in the signal, identify their modulation as being due to the (unexpectedly changing) polhode motion of the gyro and calculate the necessary rate of decrease in the polhode periodicity.

- The HF data were sufficiently detailed not just to identify the periodicities of both the main interference and the modulation that it undergoes, but to see how the modulation changes in enough detail to be able to predict, with the use of the necessary theory regarding unstable polhoding of a gyroscopic system, how it is changing during the gaps between the bursts of data (the majority of the time) and thus construct a complete evolution of the system (gyro motion).
- The available data concerning the main interference together with the modulation it undergoes were sufficient to predict the specific uneven charge distribution on the surface of the gyro that must be causing the signal with that specific signature.
- The data (for each gyro) lead to one unique combination of spin rate, spin-down rate, polhoding and TFM; that is to say, no alternative arrangement of trapped flux, or indeed any other phenomenon, could have produced the observed dataset, with its specific perturbations from what had initially been expected.

The important question in each case is: Can we use the data to distinguish between the postulated effect or phenomenon (the gyro spinning at a certain rate or polhoding in a certain way, for example) and other possible situations? I see the analysis of the counterfactual sensitivity between experimental evidence and target hypotheses combined with a probabilistic severity criterion of the type advocated by Mayo as the best way to both describe the reality of scientific experimentation in a useful way that is meaningful to scientists and reflects their concerns, and to set out clear aims and goals for test passing that reflect epistemological requirements but avoid over-simplistic or rigid criteria that scientists often seem to adopt, as reflected by the concerns expressed by NASA. I should mention again that both Woodward and Mayo insist that severity comes in degrees, I expressed this (to paraphrase Mayo) as

a probabilistic SC requiring of, a test, T, (which as we saw above is obtained for the specifics of the experiment) that:

> h is (highly) unlikely to pass T if h is false (or, h is (highly) unlikely to fail T if h is true)

Here we can see our hypothesis as the TFM, or more explicitly: the TFM is an accurate representation of the patch effects that caused the anomalies in the data. The test is the degree to which the TFM matches the data. So the key question is:

Is it possible that, had trapped flux as described by the TFM *not* been the cause of the perturbations, that the team would have been able to devise a detailed TFM that matched the data to the extreme precision that it did?

The resounding answer of the GP-B team, which appears to be rigorously supported by their results, is that no: it would have been totally impossible to devise a mapping that fitted the data as accurately as it did (and they go to great lengths to show how accurate that match is) had there been some other cause of the perturbation. So, the dataset can be used to distinguish between the use-constructed hypotheses they put forward (essentially that the TFM they devised was the cause of the anomalies) and any other postulated explanation. This could be seen as a version of the famous no miracles argument: it would take a miracle for some other cause to mirror the effects of the postulated TFM so perfectly.

The crucial claim is that had any of their hypotheses been false, then they would not have been able to come up with such a match between hypotheses and data. Thus, each of these claims is a clear case of double-counting, insofar as the same total dataset of the SQUID readout was used both to execute the TFM and to arrive at a unique distribution of electrical potential that was postulated as being responsible for the observed anomalies and perturbations, and to confirm that model as being the only possible arrangement that could have been produced that motion, since the model predicts the gyro motion actually observed. So, we are using the same dataset both to construct our hypothesis or model and then to test it and decide that we have got it right. However, by making the counterfactual sensitivity explicit and considering

the test that the corresponding hypothesis has been exposed to, we can indeed see them as being stringent and severe tests. Combining the 4 (or 5) major claims that I have just listed, we can arrive at the following counterfactual statements:

- Had the source of the electromagnetic fluctuations not been trapped or "frozen" in the gyro surface, it would not have yielded a signal with a periodicity equal to the gyro spin rate. (Indeed, had the accumulation of flux been free to move, we would have expected its effects to dissipate as the experiment advanced and it smoothed its distribution over the surface. Certainly it would not to have provided additional information on gyro motion.)
- Had an effect other than varying polhode phase and period been attenuating the gyro motion, it would not have shown the characteristic jumps (in 2 of the gyros) and monotonic damping that was observed; neither would there have been the shifts in orientation that coincided with resonance between extremely high harmonics of the (changing) polhode period and the spacecraft roll period.
- Had any other distribution of flux been trapped in the surface of each gyro, it would not have produced these exact, reproducible and predictable perturbations.

The fluctuations in the HF signal clearly show a periodicity that is extremely close to the spin frequency of the gyros, as measured during the spin-up and in-orbit initiation period. Given that the source of the perturbations was known to be electromagnetic (from the post-Science Phase calibration data) once the periodicity was established as that of the gyro spin rate, then it seems impossible to deny that the effect is due to charge, or magnetic flux, rotating with the gyros, that is, frozen in the surface. This would seem to entirely vindicate the first of the 3 counterfactual statements above.

The study of the attenuation of this perturbation, with the calculation of the changing periodicity and the observation of major jumps early on in the motion of 2 of the 4 gyros, likewise seems to attest to the truth of the second counterfactual statement above. Although the mechanism of energy

dissipation from the spinning rotors is not understood exactly (it has been suggested that this may involve internal stresses and strains within the body of the rotor), the migration of polhode motion to the principal axis with the largest moment of inertia and the damping of the motion around this axis are predicted from both theoretical and empirical results, and match exactly the modulation observed.

It is the third of the 3 counterfactual statements above that is most controversial and the justification for it is the nearly 6 years of data processing work that the team carried out after the mission had been completed. It would be impossible to give the details here, suffice it to say that the trapped flux maps that the team drew up were arrived at by 2 independent methods which yielded essentially the same results.

So, let me return to Mayo's SC and try to establish whether this, which may be considered to be a case of double-counting, can nonetheless be considered to be a severe test of the hypotheses behind the counterfactuals. Mayo talks much of the New Experimentalists whose, "experimental narratives offer a rich source from which to extricate how reliable data are obtained and used to learn about experimental processes." (Mayo, 1996, p. 58) Mayo joins the new experimentalists in "rejecting old-style accounts of confirmation as the wrong way to go" and seems to join with John Norton in telling us that: "The complexities and context dependencies of actual scientific practice just seem recalcitrant to the kind of uniform treatment dreamt of by philosophers of induction." (Mayo, 1996, p. 67) Her argument hinges on the fact that "the designing, modelling and analysing of experiments [are] activities that receive structure by means of statistical methods and activities." (Mayo, 1996, p. 58) The key to this stringency can be seen to lie in the question: Can we use the data to distinguish between true and false hypotheses? Or alternatively: Could we expect the data and the (use-constructed) hypothesis to match to the extent that they do if the hypotheses were false? As performed by the GP-B team in a way that leads to severe testing of their hypotheses by the total dataset that they used to construct those same hypotheses. It is this SC which I believe can be deployed to explain and justify the empirical warrant of the GP-B team's claims.

GP-B is a perfect example of the lessons Mayo points to but laments have not been fully taken to heart:

> Actual experimental inquiries . . . focus on manifold local tasks: checking instruments, ruling out extraneous factors, getting accuracy estimates, distinguishing real effect from artefact, and estimating the effects of background facts.
>
> (Mayo, 1996, p. xiii)

She accurately and succinctly captures the actual work of scientists in the field when they apply all their weaponry to solve the problems that experiments throw up and the GP-B team are an example of her

> shrewd inquisitors of errors, [who] interact with them, simulate them (with models and computers), amplify them. . . make them talk.
>
> (Mayo, 1996, p. 4)

6.4 Society's Tightest Demands

6.4.1 Criminal law

So far in this chapter, I have considered two different approaches to the analysis of the GP-B results, in the light of early criticism that the data processing received and particularly assessing whether accusations of double-counting are damning (or whether they should be, in this case). First, in Section 6.2, via a Quinean-like WoK, I adopted what I like to think of as being the most theoretical of the 3 approaches. Then, in Section 6.3, I brought together the work of Bogen and Woodward on the one hand, and Mayo on the other, in what I see as a more practical or pragmatic approach; one that I hope is in keeping with the ideas of New Experimentalism. My aim there was to show how experimentation really does take on "a life of its own," in the celebrated 1983 phrase of Ian Hacking's, and experimenters move forward towards the generation of novel knowledge through demonstrating flexibility and adaptability in their attempts to identify genuine difference-makers. Now, in this section, I want to examine the GP-B results and findings (that is, both the validity of the data

analysis methods and the claim to have thereby confirmed the predictions of GTR) in what I see as a more social arena. For this analysis, I will consider GP-B to be an undisputed instance of double-counting, or what in Chapter 4 I explain is also known as an example of a use-constructed hypothesis, and I want to consider whether the individual members of society, and society as a whole—ultimately responsible for footing the 1,000-odd million dollar bill for GP-B—should accept the findings as valid, even if they are a clear case of such double-counting.

To do this, I will consider what our current (modern, developed, Western) society takes to be the gold standard for matters of evidence that affect it most intimately: criminal law. By considering the interface between science and criminal law, forensics, I will then attempt to judge the GP-B claims as I believe they would be judged in a criminal court; as if the GP-B claims were the prosecution case trying to convince a sceptical (in terms of requiring considerable proof) judge: society as a whole. Although clearly the rules and requirements of evidence that apply in a laboratory and in any natural science research institute are different from those adopted in a criminal court, my aim here is—as I have said—to introduce a more social angle into my analysis. I thereby hope to add one more string, as it were, to the bow I am using to shoot probing arrows into the entire episode that GP-B represents of the generation of novel knowledge through scientific experimentation.

Society cannot be expected necessarily to adopt the same standards as theoretical physicists or epistemologists, but it is important for interested lay members of the public to arrive at informed opinions regarding the scientific activity their society undertakes through the practices and research projects of its scientists. If the public is to move away from the extremes of blind reliance

on expertise[9] and the sensational effects of brash media stunts,[10], then it seems correct and reasonable to bring into play the standards that society adopts in

[9]In his recent book (Nieto-Galan, 2016) reviewing the history of interactions and relationships between public lay knowledge and expertise by looking at different aspects of the presentation of science to the public or the different arenas and methods used for this, the last chapter, titled "Democratic science", is where the author Nieto-Galan considers the most recent developments in the area and introduces the idea of "the participatory turn". This term he attributes to Sheila Jasanoff who in her 2003 paper laments the disconnection between the lay public on the one hand and decision making regarding science and technology, and especially the management of risk, on the other. The common theme that is echoed throughout these and many other publications dating between the two is the need for the lay public to be more involved in science policy and management, and to move away from the outdated deficit model in which knowledge is passed on in a strictly top-down fashion from all-knowing experts to the passive unquestioning receptive public.

[10]All too often, politicians, particularly, attempt to win over public opinion by appealing to the more base gut reactions, so to speak, of the public, rather than through taking the time and trouble necessary to relate a full exposition of a subject. I am thinking of examples such as the infamous incident in 1990 when the British Minister for Agriculture, Fisheries and Food, John Gummer, posed for press photographers while (supposedly) feeding a hamburger to his 4-year-old daughter. The stunt was an attempt to convince the public that it was perfectly safe to eat British beef, at the height of the BSE or "mad cow" crisis in Britain. In her analysis of this media stunt in the chapter titled "Civil Epistemology" of her book *Designs on Nature: Science and Democracy in Europe and the United States*, Jasanoff, 2005, p. 256, tells us: "It was an act designed to meld together two age-old repertoires of trust: a father feeding his child and a state, in loco parentis, reassuring its citizens. But the performance backfired, and both the performance and the manner of its backfiring offer illuminating insights." In the uproar that followed the incident, it was then claimed that the 4-year-old minister's daughter eating the hamburger had been staged and in fact a civil servant had previously removed a bite from the hamburger. It is interesting to note that even today the Spanish blood donation service refuses blood from anyone who spent a total of 12 months or more in the UK between 1980 and 1996, due to the risk of harbouring latent CJD infection (`http://www.donarsangre.org/puedo-donar-si/` accessed on 04/08/2016).Indeed, in Spain, in a much earlier media stunt back in 1966 (the earliest use of television for this kind of stunt I know of in Europe), the Minister for Information and Tourism in the regime of the Franco dictatorship, Manuel Fraga (later to become Spain's last ambassador to the UK under Franco) was famously filmed bathing in the Mediterranean Sea with the US ambassador to Spain, Angier Biddle Duke, after a US B-52 bomber had exploded over the coast of Spain and shed its load of four H-bombs (some reports say 5; with the 5th never having been recovered). In his account of the whole nuclear accident, the ensuing clean-up and the media operation, the Canadian historian David Stiles tells us: "The swimming demonstration was judged by all involved to be a great success and earned the positive, front-page

other fields. Indeed, in a society where the most heinous offences are tried by a jury of (lay) peers of the accused, this seems to me to be a most fitting parallel to adopt: interested and motivated lay members of the public should be able to understand and access this standard of proof, which they may be required to put into action at any time if called upon to exercise jury service. That is why I feel it is perfectly justified to take these, the most stringent of society's norms for acceptance in matters lay individuals may be required to judge and to apply them here.

6.4.2 Reasonable doubt

Criminal law is one of the key instruments that modern society uses to protect itself; possibly ironically, from itself! (Or maybe it is more accurate to say, from antisocial elements within itself.) These two characteristics, that of providing vital protection and that of being a weapon that may at any moment be wielded against any one of us, mean that we are as strict as we can be in our demands regarding the validity of evidence in criminal law. Famously, in the Anglo-Saxon tradition of adversarial legal systems, the standard of evidence in criminal proceedings requires that a matter be proven "beyond reasonable doubt".[11] The certainty that this term aims to embody is one which rules out any other reasonable alternative.[12] In this, I see it as strictly akin to

press coverage that Duke had sought" (for a full account see Stiles, 2006, p. 62). Three of the bombs landed in southern Spain, 2 exploding on impact and dispersing plutonium over the area, while the other was lost in the sea and became known as the lost H-bomb. It was eventually recovered four months later; but the clean-up of the radioactive contamination is still a controversial subject.

[11]For example, on their webpage, the introduction to the research on this subject performed at the Institute of Criminology of the University of Cambridge tells us that: "Beyond reasonable doubt (BRD) is the standard of proof used to convict defendants charged with crimes in the English criminal justice system. If the decision maker perceives that the probability the defendant committed the crime as charged (based on the evidence) is equal or greater than their interpretation of BRD, than he/she will decide to convict. Otherwise, the decision maker will acquit the defendant." (http://www.crim.cam.ac.uk/research/beyond_reasonable_doubt/ Accessed 3/8/2106)

[12]The wording itself has been much discussed over the years and repeatedly it has been stated that it is the spirit and not the wording that is important.

Mayo's SC; I will return to this below. The history of this legal requirement, introduced into English law in the late 18th century, can be traced back to the groundbreaking work of Bacon and Locke in moving away from the medieval idea that some external otherworldly force ("God") acted through and on people, to seeing people as rational agents capable of judging the truth of matters (and also on some or other correspondence theory of truth: true statements, for example, correspond to facts pertaining to certain aspects of an objective external world that is independent of human observation). Being beyond reasonable doubt is the standard that society adopted in this context and has stuck with to protect itself for hundreds of years. It forms one of the cornerstones upon which our modern personal freedom is based. When considering this standard, in their chapter on evidence in the comprehensive "Companion to Philosophy of Law and Legal Theory" Jackson and Doran, 1999, remind us:

> it may not be possible to prove the material facts to a degree of absolute certainty. The decision will then have to be made under conditions of uncertainty. The rationalist tradition assumes that knowledge is a matter of probability and not certainty ... there is therefore a need for rules to determine what standard of proof is necessary to enable the material facts to be considered proved or not proved. The standard required must be a degree of probability ... The state must therefore bear the burden of proving guilt beyond reasonable doubt, but it is worth noting that the standard is not stretched to one of beyond all doubt.
>
> (Jackson and Doran, 1999, p. 180)

Now, in contemporary society—and far more closely related to my interests here—the evidence which is under consideration in a criminal court is often by its nature forensic, defined in the on-line version of the OED as: "the application of scientific methods and techniques to the investigation of crime".[13] In forensics, it is very common to start with the evidence and to use this evidence to construct a hypothesis. The support that we than have for the hypothesis is the very same evidence on which that hypothesis was built. This is a text-book case of a use-constructed hypothesis or double-counting. I will look at a very

[13] https://en.oxforddictionaries.com/definition/forensic accessed on 07/09/2016

common example from contemporary criminal proceedings that should be familiar to us all due to its frequent portrayal in popular contemporary culture involving detective work and criminal trials: the case of a lethal injury.

In such a case, by studying the particular structure of the wound, it may be possible to arrive at hypotheses regarding the object that was used to inflict it. For example, that it is a blunt force trauma produced by a cylindrical object with a certain diameter and range of total weights or densities, and that it was curved with a certain radius of curvature at the part that produced the impact. This could lead us to rule out the majority of possible implements as the weapon involved, and leave a very limited set of objects that fit these criteria. This may lead us to suspect that the wound was caused by a specific object—a certain tool or part of an ornament whose structure matches the characteristics we have determined, to an extremely high degree—and to hypothesise that this was the object used to commit the crime (and maybe due to many other considerations of timing and location; and of course of additional evidence, such as DNA traces, for example). Since I am interested in the parallel here with procedures in natural science, for my purposes I feel it is best and perfectly legitimate to ignore the last step of tying the weapon to an agent responsible for the action. Instead I consider only the finding that the object in question is the murder weapon. The evidence in favour of such a hypothesis is a mixture of the evidence that we used to construct the hypothesis in the first place, that is the wound itself, and all our background knowledge together with a whole collection of *ceteris paribus* clauses. It is not the case that such hypotheses are deemed untested and not admitted in court on the basis that they could not have failed to fit the evidence (wound) since they were built using that very same evidence (and precisely to fit it as closely as possible). Indeed, it would seem quite ridiculous in court proceedings to call into question the validity of such hypotheses because they could not have failed to match. Instead, they are only doubted if some other, alternative explanation for the results (the lethal wound), such as a different object with similar characteristics, for example, can be found and brought into evidence. In such a case, that the hypothesis matches the evidence is as stringent or severe a test as we can require; there is no certainly, but our use-constructed hypothesis has certainly passed an extremely stringent test.

Let me just return here to the importance that Mayo gives, in her 2008 paper (Mayo, 2008), to the procedure by which a non-UN hypothesis is reached. That paper was written in response to criticism of SC (and particularly in response to the criticism that double-counting is an all or nothing affair which leads to a dilemma that Mayo sees as completely false). The criticism is that if we allow double-counting, then every such use-constructed hypothesis cannot fail to match the data and so it is minimally severe (totally "insevere") in that whatever the data, we get a hypothesis that is supported by them. However, as Mayo goes to great lengths to stress, her SC is not limited to the idea that data must fit a hypothesis to a suitable degree; it is the likelihood of a use-construction procedure producing a hypothesis (that by definition fits the data) even though the hypothesis *is false* that determines severity in these cases. The procedure or construction rule, R as Mayo denotes it, is explicitly assigned importance; and we need R to be reliable in order to come up with severe tests, as in this example of forensics: the hypotheses that we constructed with the observed data (characteristics of the wound) in mind are reliably linked to the conclusion (that a specific type of object was used to inflict the wound) by means of using procedures to construct the hypotheses that are known to be reliable. We therefore must want to see the matches as severe tests. Denoting the use-constructed hypothesis as $H(x_0)$, Mayo tells us that:

> ...the inference to $H(x_0)$ must be evaluated in relation to the construction rule actually used. ...we evaluate the severity with which $H(x_0)$ has passed by considering the stringency of the rule R by which it was constructed, taking account of the particular data achieved.
>
> (Mayo, 2008, pp. 874 and 877)

Our construction rule in this case involves all our previous experience with impact wounds and possibly experiments performed with the suspected weapon. These are procedures which have been shown time and again to be reliable and therefore we have a high degree of expectation that in this case they are also leading reliably from the evidence to the hypothesis and the fact that the two fit is indeed all the "proof" that is needed.

Finally, I just want to add a note on the qualitative, and in that sense vague, nature of severity. Mayo, once again points out that:

> ...there is no suggestion that one can always calculate the severity associated with a use-construction rule ... the onus is on the tester to show the hypothesis in question has been well tested.
> (Mayo, 2008, p 874)

This quotation shouts to me of its similarity with criminal court proceedings and the onus of the standard of proof, particularly in relation to the quotation above from Jackson and Doran, and therefore seems particularly pertinent to me here.

6.4.3 Reliable experimental procedures

I hope the analogy I am trying to draw is clear. In the case of GP-B—if we accept, as I stipulate at the start of this section, the claim that it is a case of double-counting: the dataset that was used to arrive at the extremely complicated data model and TFM that allowed the team to separate out the unwanted effects, is the very same dataset used to show that the data model and hence TFM are correct—the use of double-counting in no way means that the test to which the hypothesis is subjected is not severe. The global hypothesis that I am interested in here is that the undesired effects caused by the patch effects on the gyros can be removed from the original noisy dataset to leave "clean" post-processing data that contains the data that are the evidence for the effects predicted by GTR. This was achieved via the TFM (performed on each of the 4 gyroscopes) using the noisy data. This is analogous to the claim that by studying a lethal wound—noisy dataset—we can extract certain characteristics of the weapon that must have caused it—the specific characteristics of the underlying electromagnetic patches on each gyro. We can then be confident that we have the correct TFM because, from the characteristics that we extracted from the noisy data (specifically the changing polhode characteristics and the classical torques that were determined to be acting on the gyros—equivalent to the certain diameter and range of total weights or densities together with the radius of curvature of the weapon) the

team then constructed the only TFM that would have led to precisely those parameters—our murder weapon.

Furthermore, just as in the criminal case, we are open to alternative explanations. Indeed, a considerable part of the analysis of the GP-B data involved assessing just whether any other distribution of trapped flux would have produced similar effects.[14] That was achieved through the many different layers of modelling leading on one unique TFM for each gyro. Here the team brought into play all their additional background knowledge and applied their immense experience in many different fields of physics to the problem at hand. Once again, we can see this as analogous to the work performed in the forensic laboratories to arrive at the final characteristics of the murder weapon and quite possibly in testing candidates in the laboratory (as the GP-B team did with the flight-standard gyro rotors they had in their own laboratories).

Now, what of the procedures used and the reliability of the construction rule, R? This is where the redundancy built into GP-B via the use of 4 independent gyros comes into play. Despite the team's confidence in the physics they built their procedure on, it was the first time such TFM analysis had been performed. So there are two aspects to consider here. First of all, the TFM procedure yielded wholly consistent results for all four gyros completely independently. This builds confidence massively in R: the way they moved from the data to the characteristics of the trapped flux. Whereas it may have been argued that the procedure was totally ad hoc, bound to succeed and a blatant case of over-fitting if the method had only been applied to one instance of a gyro,[15] the fact that the same method gave meaningful and consistent results on four separate occasions cannot be written off in the same way. The other aspect is that the severity of the test is not derived merely from the fit between data and hypothesis (which of course is guaranteed in cases of

[14]Bear in mind that I have never questioned the electromagnetic nature of the origin of the interference as that was established using additional data collected during the post-science calibration phase, not using the noisy dataset acquired during the Science Phase of the mission.

[15]In fact, if patch effects had only been detected on one gyro, the only possible action would have been to discard the data from that gyro, as the procedure to clean-up the data would have been almost impossible to justify and could always have been the result of gross over-fitting.

double-counting) but, as I say above and Mayo insists, it is the likelihood of a use-construction procedure producing a hypothesis that passes the test in question even though the hypothesis is false that determines severity in these cases. So we have to ask, as the jury is asked to ponder in criminal proceedings when no specific alternative has been put forward: Is it possible that the TFM was incorrect and a mere invention of the GP-B team, and yet that it fits the data so well? The analogous question here in the murder trial would be whether, given that only one item has been identified that fits all the characteristics of the fatal wound, is it possible that in fact this is just a coincidence and that there is another cause that remains undetected. I have no doubt that for GP-B, the case has been proved beyond reasonable doubt.

So, my argument in this section is both simple and straightforward. If the GP-B findings and therefore the overall claim of confirming GTR are accused of being an instance of double-counting, and that is certainly a criticism which may very reasonably be levelled against them, and if they are to be judged by society in general (by lay members of the public), then by applying the standards that society demands of evidence and findings in the most critical and exacting cases that it faces in everyday life and where lay members of the public are called upon to make a decision, that is, the burden of proof in criminal proceedings, then the fact that GP-B is a case of double-counting cannot be seen as a damning characteristic or indeed to call its findings into question as this is precisely the nature of much of the evidence—forensic evidence—that is held up as decisive and capable of being beyond reasonable doubt in criminal courts every day.

Chapter 7

Closing Remarks and Further Research

Such conclusions as I have arrived at have all been made; I do not consider it necessary or fitting to recount them here. Like so much of philosophy, and indeed of life, it is the journey that is the important part of this work, and not the final destination. My objective here has certainly been to examine GP-B as a case study of the generation of knowledge through scientific experimentation, and to pass my informed judgement on the processes and procedures undertaken as part of it. That judgement, however, will certainly not change what has already occurred, and my hope here must be to influence the future. Once again this points to the importance of the critical process, more than to the specific object of my critical analysis. It is my hope that this work will serve me and others as a point of departure and as a tool to be used in the analysis and criticism of similar episodes, both in experimentation in fundamental physics, and in other fields. It has certainly formed a solid base from which I feel I can launch myself into similar tasks of critical analysis.

Having taken on board so much physics for this work, which I thought I had left behind me years before embarking on this project, I am now keen to continue down the path that the philosophy of physics opens up before me. The study of General Relativity was entirely new to me when I started this work. It has proved to be both fascinating and a most fruitful mixture of physics and philosophy. I am eager to continue down this path and to deepen my knowledge both of Einstein's theory and of the broader questions

concerning the nature of space and time that it throws up at us. Immediately on considering these questions I find myself also draw into the very human realm of perception: the perception of time. The work that I have been exposed to in this field has more than whetted my appetite to delve into the complicated study of the connections between our perception of time and the concept of time within physics. The work of David Eagelman that I was fortunate enough to see him present here in Barcelona a few years ago has left an important mark, and is one of the areas that I will continue to read on and hope to become more involved in.

Through my work with the GRECC research Group, I have already undertaken some work on perception and the philosophy of mind: this is a more general area leading on from the perception of time that also holds great fascination for me. Maybe due to my background in physics and my (limited) experience in the philosophy of science, the interface between consciousness and technology is an area that I have taken an interest in and hope to continue to study. The ideas related to the fields of augmented and enhanced reality, and the possibilities of human physical enhancement that technology is bringing into our grasp are areas where I am sure philosophers will have much to contribute and it is an area I am also keen to work in. I have already undertaken work within the field of technoscience and so feel this is a natural route for me to follow further.

Another of the aspects of the research for this thesis that I have been active in is related to the interdisciplinary encounter of the history and philosophy of science. As a member of the Barcelona History and Philosophy of Science Research Group, again this is a field that I am already active in and intend to continue my work in. The work on GP-B has certainly lent itself extremely well to this area and offers me the opportunity of joint research with historians of science. It brings together very neatly many aspects of the research we have embarked on into both scientific rationality and scientific discovery.

With so many options and avenues open to me, and feeling drawn towards many of them, there can be no doubt that the research behind the production of this thesis will serve as an invaluable base from which to expand. As one of my heroes might say at this point: “To infinity, and beyond!”.

Bibliography

Abbott, B. P. et al. (2016). "Observation of Gravitational Waves from a Binary Black Hole Merger". In: *Phys. Rev. Lett.* 116 (6), p. 061102. DOI: 10.1103/PhysRevLett.116.061102. URL: http://link.aps.org/doi/10.1103/PhysRevLett.116.061102.

Ackermann, Robert (1989). "The New Experimentalism". In: *British Journal for the Philosophy of Science* 40, pp. 185–190. DOI: 10.1093/bjps/40.2.185.

Adler, Ronald J. (2015). "The three-fold theoretical basis of the Gravity Probe B gyro precession calculation". In: *Classical and Quantum Gravity* 32.22, p. 224002. URL: http://stacks.iop.org/0264-9381/32/i=22/a=224002.

Balogh, A. and G. Giampieri (2002). "Mercury: the planet and its orbit". In: *Reports On Progress In Physics* 65, 529–560.

Bencze, W. J., R. W. Brumley, M. L. Eglington, D. N. Hipkins, T. J. Holmes, B. W. Parkinson, Y. Ohshima, and C. W. F. Everitt (2015). "The Gravity Probe B electrostatic gyroscope suspension system (GSS)". In: *Classical and Quantum Gravity* 32.22, p. 224005. URL: http://stacks.iop.org/0264-9381/32/i=22/a=224005.

Bogen, James and James Woodward (1988). "Saving the phenomena". In: *The Philosophical Review* 97, 303–352.

– (1992). "Observations, theories and the evolution of the human spirit". In: *Philosophy of Science* 59, 590–611.

– (2005). "Evading the Irs". In: *Poznan Studies in the Philosophy of the Sciences and the Humanities* 86.1, pp. 233–268.

Brans, Carl H. *The roots of scalar-tensor theory: an approximate history*. arXiv:gr-qc/0506063v1. accessed on 05/12/2005.

Buchanan, S., C. W. F. Everitt, B. Parkinson, J. P. Turneaure, and G. M. Keiser (2000). "Cryogenic Gyroscopes for the Relativity Mission". In: *Physica B* 280, 497–498.

Buchman, S., J. A. Lipa, G. M. Keiser, B. Muhlfelder, and J. P. Turneaure (2015). "The Gravity Probe B gyroscope". In: *Classical and Quantum Gravity* 32.22, p. 224004. URL: http://stacks.iop.org/0264-9381/32/i=22/a=224004.

Cannon, R. H. (1963). "International Union of Theoretical and Applied Mechanics Kreiselprobleme Gyrodynamics Symposium, Celerina, August 20-23, 1962". In: ed. by H. Ziegleb. Berlin: Springer-Verlag. Chap. "Requirements and Design for a Special Gyro for Measuring General Relativity Effects From an Astronomical Satellite", pp. 145–157.

Carlson, Matthew (2015). "Logic and the Structure of the Web of Belief". In: *Journal for the History of Analytical Philosophy* 3.5, pp. 1–26.

Cartwright, N. (1989). *Nature's Capacities and Their Measurement*. Oxford: Clarendon Press.

Ciufolini, I. and E. Pavlis (2004). "A Confirmation of the General Relativistic Prediction of the Lense-Thirring Effect". In: *Nature* 431(7011), pp. 958–960.

Ciufolini, I., E. Pavlis, F. Chieppa, E. Fernandes-Vieira, and J. Perez-Mercader (1998). "Test of General Relativity and Measurement of the Lense-Thirring Effect with Two Earth Satellites". In: *Science* 279(5359).

Collas, P. and D. Klein (2004). "Frame Dragging Anomalies for Rotating Bodies". In: *General Relativity and Gravitation* 36.5, pp. 1197–1206.

Collins, Harry M. and Trevor Pinch (1993). *The Golem: What You Should Know about Science*. Canto (Cambridge University Press). Cambridge: Cambridge University Press. ISBN: 9780521645508.

Conklin, J. W., M. I. Heifetz, T. Holmes, M. Al-Meshari, B. W. Parkinson, A. S. Silbergleit, C. W. F. Everitt, A. Al-Jaadan, G. M. Keiser, B. Muhlfelder, V. G. Solomonik, and H. Al-Jabreen (2015). "Gravity Probe B data analysis: III. Estimation tools and analysis results". In: *Classical and Quantum Gravity* 32.22, p. 224020. URL: http://stacks.iop.org/0264-9381/32/i=22/a=224020.

Cowan, C. L. Jr., F. Reines, F. B. Harrison, and A. D. Kruse H. W. McGuire (1956). "Detection of the Free Neutrino: A Confirmation". In: *Science* 124, pp. 103–104. DOI: 10.1126/science.124.3212.103.

Cutting, G. (2001). "A Companion to the Philosophy of Science". In: ed. by W. H. Newton-Smith. Blackwell Companions to Philosophy. Oxford: Blackwell. Chap. "Scientific Methodology", pp. 423 –432.

Damour, T. (1992). "General Relativity and Experiment: a brief review". In: *Classical and Quantum Gravity* 9.S, S55 –S59. URL: http://stacks.iop.org/0264-9381/9/i=S/a=017.

Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Einstein, A. (1952[1916]). "The Principle of Relativity: A Collection of Original Memoirs on the Special and General Theory of Relativity". In: ed. by H.A. Lorentz, A. Einstein, H. Minkowski, H. Weyl, and A Sommerfeld. New York: Dover. Chap. "The Foundation of the General Theory of Relativity", pp. 109 –164.

– (2005[1922]). *The Meaning of Relativity (Republished 2005 with intro. by Brian Green)*. Princeton: Princeton University Press.

Everitt, C. W. F., D. B. DeBra, P. Wiktor, J. Kasdin, and Y. Jafry (1993). "Automatic Control in Aerospace 1992". In: ed. by D. B. DeBra and E. Gottzein. Oxford - New York - Seoul - Tokyo: Pergamon Press. Chap. "Gravity Probe-B: A Gyro Test of General Relativity in a Satellite", pp. 241 –244.

Everitt, C. W. F., M. Adams, W. Bencze, S. Buchman, B. Clarke, J. W. Conklin, D. B. DeBra, M. Dolphin, M. Heifetz, D. Hipkins, T. Holmes, G. M. Keiser, J. Kolodziejczak, J. Li, J. Lipa, J. M. Lockhart, J. C. Mester, B. Muhlfelder, Y. Ohshima, B. W. Parkinson, M. Salomon, A. Silbergleit, V. Solomonik, K. Stahl, M. Taber, J. P. Turneaure, S. Wang, and P. W. Worden (2009). "Gravity Probe B Data Analysis". In: *Space Science Reviews* 148.1, pp. 53–69. ISSN: 1572-9672. DOI: 10.1007/s11214-009-9524-7. URL: http://dx.doi.org/10.1007/s11214-009-9524-7.

Everitt, C. W. F., D. B. DeBra, B. W. Parkinson, J. P. Turneaure, J. W. Conklin, M. I. Heifetz, G. M. Keiser, A. S. Silbergleit, T. Holmes, J. Kolodziejczak, M. Al-Meshari, J. C. Mester, B. Muhlfelder, V. G. Solomonik, K. Stahl, P. W. Worden, W. Bencze, S. Buchman, B. Clarke, A. Al-Jadaan, H. Al-Jibreen, J. Li, J. A. Lipa, J. M. Lockhart, B. Al-Suwaidan, M. Taber, and S. Wang (2011). "Gravity Probe B: Final Results of a Space Experiment to Test General Relativity". In: *Phys. Rev. Lett.* 106 (22), p. 221101. DOI: 10.1103/PhysRevLett.106.

221101. URL: http://link.aps.org/doi/10.1103/PhysRevLett.106.221101.

Everitt, C W F, B Muhlfelder, D B DeBra, B W Parkinson, J P Turneaure, A S Silbergleit, E B Acworth, M Adams, R Adler, W J Bencze, J E Berberian, R J Bernier, K A Bower, R W Brumley, S Buchman, K Burns, B Clarke, J W Conklin, M L Eglington, G Green, G Gutt, D H Gwo, G Hanuschak, X He, M I Heifetz, D N Hipkins, T J Holmes, R A Kahn, G M Keiser, J A Kozaczuk, T Langenstein, J Li, J A Lipa, J M Lockhart, M Luo, I Mandel, F Marcelja, J C Mester, A Ndili, Y Ohshima, J Overduin, M Salomon, D I Santiago, P Shestople, V G Solomonik, K Stahl, M Taber, R A Van Patten, S Wang, J R Wade, P W Worden Jr, N Bartel, L Herman, D E Lebach, M Ratner, R R Ransom, I I Shapiro, H Small, B Stroozas, R Geveden, J H Goebel, J Horack, J Kolodziejczak, A J Lyons, J Olivier, P Peters, M Smith, W Till, L Wooten, W Reeve, M Anderson, N R Bennett, K Burns, H Dougherty, P Dulgov, D Frank, L W Huff, R Katz, J Kirschenbaum, G Mason, D Murray, R Parmley, M I Ratner, G Reynolds, P Rittmuller, P F Schweiger, S Shehata, K Triebes, J VandenBeukel, R Vassar, T Al-Saud, A Al-Jadaan, H Al-Jibreen, M Al-Meshari, and B Al-Suwaidan (2015). "The Gravity Probe B test of general relativity". In: *Classical and Quantum Gravity* 32.22, p. 224001. URL: http://stacks.iop.org/0264-9381/32/i=22/a=224001.

Fairbank, J. D., B. S. Jr Deaver, C. W. F. Everitt, and P. F. Michelson (1988). *Near Zero: New Frontiers of Physics*. New York: W H Freeman and Co.

Friedman, M. (1999). *Reconsidering Logical Positivism*. Cambridge: Cambridge University Press. ISBN: 9780521624763.

Giere, Ronald N. (1988). *Explaining Science: A cognitive approach*. Chicago: The University of Chicago Press.

Goldman, Alvin (1986). *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.

GP-B Mission Update: 21/03/2008. https://einstein.stanford.edu/highlights/hl_032108.html. accessed on 09/08/2016.

GP-B Mission Update: 26/09/2008. https://einstein.stanford.edu/highlights/hl_092608.html. accessed on 09/08/2016.

Haack, Susan (2003). *Defending Science-within Reason: Between Scientism and Cynicism*. Amherst, New York: Prometheus Books. ISBN: 9781591021179.

Heifetz, M. I., G. M. Keiser, A. S. Krechetof, and A. S. Silbergleit (2000). "FUSION 2000 Proceedings of the Third International Conference on Information Fusion July 10-13, 2000, Cité Des Sciences Et de L'Industrie, Paris, France". In: vol. 2. International Society of Information Fusion. Chap. "Multisensor Data Integration in the NASA/Stanford Gravity Probe B Relativity Mission", WEC5–WEC16.

Hitchcock, Christopher and Elliott Sober (2004). "Prediction Versus Accommodation and the Risk of Overfitting". In: *The British Journal for the Philosophy of Science* 55.1, pp. 1–34. DOI: 10.1093/bjps/55.1.1. URL: http://bjps.oxfordjournals.org/content/55/1/1.abstract.

Hoefer, Carl (1994). "Einstein's struggle for a Machian gravitation theory". In: *Studies in the History and Philosophy of Science* 25.3, pp. 287 –335.

Hylton, Peter (2014). "Willard van Orman Quine". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2014.

Jackson, John and Sean Doran (1999). "Companion to Philosophy of Law and Legal Theory". In: ed. by Dennis Patterson. Wiley-Blackwell. Chap. Evidence.

Janssen, Michel and Christoph Lehner, eds. (2014). *The Cambridge Companion to Einstein*. Cambridge: Cambridge University Press. ISBN: 9781139024525. DOI: 10.1017/CCO9781139024525. URL: https://www.cambridge.org/core/books/the-cambridge-companion-to-einstein/130837F4D6A9D9BA84F928AB82EDF692.

Jasanoff, S. (2005). *Designs on Nature: Science and Democracy in Europe and the United States*. Princeton paperbacks. Princeton University Press. ISBN: 9780691118116.

Kahn, Robert (2007). *Gravity Probe B–Post Flight Analysis · Final Report*. Tech. rep. Stanford University, p. 616. URL: http://einstein.stanford.edu/content/final_report/GPB_FinalPFAR-091907-prnt.pdf.

– (2008). *Gravity Probe B Science Results–NASA Final Report*. Tech. rep. Stanford University, p. 84. URL: https://einstein.stanford.edu/content/final_report/GPB_Final_NASA_Report-020509-web.pdf.

Kargon, Robert, Stuart W. Leslie, and Erica Schoenberger (1992). "Big Science: The growth of large-scale research". In: ed. by Peter Galison and Bruce

Hevly. Chap. "Far Beyond Big Science: Science Regions and the Organization of Research and Development".

Keiser, G. M. (2003). "General Relativity Experiments In Space". In: *Advanced Space Research* 32.

Keiser, G. M., J. Kolodziejczak, and A. S. Silbergleit (2009). "Misalignment and Resonance Torques and Their Treatment in the GP-B Data Analysis". In: *Space Science Reviews* 148.1, pp. 383–395. ISSN: 1572-9672. DOI: 10.1007/s11214-009-9516-7. URL: http://dx.doi.org/10.1007/s11214-009-9516-7.

Kuhn, T.S. (1996[1962]). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lammerzahl, C., G. Ahlers, N. Ashby, M. Barmatz, P. L. Biermann, H. Dittus, V. Dohm, R. Duncan, K. Gibble, J. Lipa, N. Lockerbie, N. Mulders, and C. Salomon (2004). "Experiments in fundamental physics scheduled and in development for the ISS". In: *General Relativity and Gravitation* 36, pp. 615–649. DOI: 10.1023/B:GERG.0000010734.62571.b4.

Li, J., G. M. Keiser, J. M. Lockhart, Y. Ohshima, and P. Shestople (2015). "Timing system design and tests for the Gravity Probe B relativity mission". In: *Classical and Quantum Gravity* 32.22, p. 224014. URL: http://stacks.iop.org/0264-9381/32/i=22/a=224014.

Mattingly, D. (2005). "Modern Tests of Lorentz Invariance". In: *Living Reviews in Relativity* 8.5. DOI: 10.12942/lrr-2005-5. URL: http://relativity.livingreviews.org/Articles/lrr-2005-5/.

Mayo, Deborah G. (1991). "Novel Evidence and Severe Tests". In: *Philosophy of Science* 58, pp. 523–552.

– (1996). *Error and the Growth of Experimental Knowledge*. Science and Its Conceptual Foundations series. University of Chicago Press. ISBN: 9780226511986.

– (2008). "How to Discount Double-Counting When It Counts: Some Clarifications". In: *British Journal for Philosophy of Science* 59, pp. 857–879.

Merton, Robert K. ((1973)[1942]). "The Sociology of Science: Theoretical and Empirical Investigations". In: Chicago: University of Chicago Press. Chap. "The Normative Structure of Science". ISBN: 978-0-226-52091-9.

Misner, C. W., K. S. Thorne, and J. A. Wheeler (1973). *Gravitation*. Gravitation parte 3. W. H. Freeman. ISBN: 9780716703440.

Morgan, M. S. and M Morrison (1999). "Models as Mediators: Perspectives on Natural and Social Scienc". In: ed. by M. S. Morgan and M Morrison. Cambridge: Cambridge University Press. Chap. "Models as Mediating Instruments", pp. 10 –37.

Muhlfelder, B., J. Lockhart, H. Aljabreen, B. Clarke, G. Gutt, and M. Luo (2015). "Gravity Probe B gyroscope readout system". In: *Classical and Quantum Gravity* 32.22, p. 224006. URL: http://stacks.iop.org/0264-9381/32/i=22/a=224006.

NASA (2008). *Report of the 2008 Senior Review of the Astrophysics Division Operating Missions April 22 – 25, 2008*. Tech. rep. National Aeronautics and Space Administration, p. 27. URL: http://science.nasa.gov/media/medialibrary/2010/03/31/2008_Astro_SR_20Final_Report.pdf.

Nieto-Galan, Agusti (2016). *Science in the public sphere: a history of lay knowledge and expertise*. Hoboken, NJ: Routledge.

Nordtvedt, Kenneth (1968). "Equivalence Principle for Massive Bodies. I. Phenomenology and II. Theory". In: *Physics Review* 169, pp. 1014–1025.

Norton, John D. (2005). "The Genesis Of General Relativity Volume 3: Theories of Gravitation on the Twilight of Classical Physics. Part I". In: ed. by Jürgen Renn. Norwell: Kluwer Academic Publishers. Chap. "Einstein, Nordström and the Early Demise of Scalar, Lorentz Covariant Theories of Gravitation".

– (2010). "Symmetries in Physics: Philosophical Reflections". In: ed. by K. Brading and E. Castellani. Cambridge: Cambridge University Press. Chap. "General Covariance, Gauge Theories and the Kretschmann Objection".

O'Connell, R. F. (1968). "Schiff's Proposed Gyroscope Experiment as a Test of the Scalar-Tensor Theory of General Relativity". In: *Physical Review Letters* 20, pp. 69–71. DOI: 10.1103/PhysRevLett.20.69.

Overduin, J. M. (2015). "Spacetime, spin and Gravity Probe B". In: *Classical and Quantum Gravity* 32.22, p. 224003. URL: http://stacks.iop.org/0264-9381/32/i=22/a=224003.

Oxford English Dictionary. http://www.oxfordlearnersdictionaries.com/definition/english/trouble-at-t-mill. accessed on 13/09/2016.

Penrose, R. (2004). *The Road to Reality: A Complete Guide to the Laws of the Universe*. London: Jonathan Cape.

Perimeter Institute for Theoretical Physics. https://www.perimeterinstitute.ca/pi-kids-are-asking-how-does-egg-stand-its-tip. accessed on 19/06/2016.

Pugh, George E. (2003[1959]). "Nonlinear Gravitodynamics, The Lense-Thirring Effect: A documentary introduction to current research". In: ed. by R. J. Ruffini and C. Sigismondi. Singapore: World Scientific Publishing. Chap. "Proposal for a Satellite Test of the Coriolis Prediction of General Relativity: WSEG Research Memorandum Number 11, November 12, 1959", pp. 414–426. DOI: 10.1142/9789812564818_0034.

Quine, W. V. (1951). "Main Trends in Recent Philosophy: Two Dogmas of Empiricism". In: *The Philosophical Review* 60.1, pp. 20–43. ISSN: 00318108, 15581470.

Reichenbach, Hans (1958). *The Philosophy of Space and Time*. Dover books explaining Science and Mathematics. New York: Dover Publications. ISBN: 9780486604435.

Schiff, L. I. (1960). "Possible New Experimental Test of General Relativity Theory". In: *Phys. Rev. Lett.* 4 (5), pp. 215–217. DOI: 10.1103/PhysRevLett.4.215. URL: http://link.aps.org/doi/10.1103/PhysRevLett.4.215.

Schilpp, Paul Arthur and Albert Einstein (1998[1949]). *Albert Einstein: Philosopher-Scientist*. The Library of Living Philosophers. Tudor Publishing Company.

Schutz, B.F. (1985). *A First Course in General Relativity*. Series in physics. Cambridge: Cambridge University Press. ISBN: 9780521277037.

Silbergleit, A., J. Conklin, D. DeBra, M. Dolphin, G. Keiser, J. Kozaczuk, D. Santiago, M. Salomon, and P. Worden (2009). "Polhode Motion, Trapped Flux, and the GP-B Science Data Analysis". In: *Space Science Reviews* 148.1, pp. 397–409. ISSN: 1572-9672. DOI: 10.1007/s11214-009-9548-z. URL: http://dx.doi.org/10.1007/s11214-009-9548-z.

Silbergleit, A. S., G. M. Keiser, J. P. Turneaure, J. W. Conklin, C. W. F. Everitt, M. I. Heifetz, T. Holmes, and P. W. Jr. Worden (2015a). "Gravity Probe B data analysis: I. Coordinate frames and analysis models". In: *Classical and*

Quantum Gravity 32.22, p. 224018. URL: http://stacks.iop.org/0264-9381/32/i=22/a=224018.

Silbergleit, A. S., J. W. Conklin, M. I. Heifetz, T. Holmes, J. Li, I. Mandel, V. G. Solomonik, K. Stahl, P. W. Jr. Worden, C. W. F. Everitt, M. Adams, J. E. Berberian, W. Bencze, B. Clarke, A. Al-Jadaan, G. M. Keiser, J. A. Kozaczuk, M. Al-Meshari, B. Muhlfelder, M. Salomon, D. I. Santiago, B. Al-Suwaidan, J. P. Turneaure, and J. Wade (2015b). "Gravity Probe B data analysis: II. Science data and their handling prior to the final analysis". In: *Classical and Quantum Gravity* 32.22, p. 224019. URL: http://stacks.iop.org/0264-9381/32/i=22/a=224019.

Sponsell, A. (2002). "Constructing a "revolution in science": the campaign to promote a favourable reception for the 1919 solar eclipse experiments". In: *British Journal for the History of Science* 35, pp. 439–467.

Stachel, J. (2001). *Einstein from 'B' to 'Z'*. Einstein Studies. Birkhäuser Boston. ISBN: 9780817641436.

Stiles, David (2006). "A Fusion Bomb over Andalucía: U.S. Information Policy and the 1966 Palomares Incident". In: *Journal of Cold War Studies* 8.8, 49–67.

Suárez, Mauricio (1999). "Models as Mediators: Perspectives on Natural and Social Scienc". In: ed. by M. S. Morgan and M Morrison. Cambridge: Cambridge University Press. Chap. "The Role of Models in the Application of Scientific Theories: Epistemological Implications".

Suppe, F. (1989). *The Semantic Conception of Theories and Scientific Realism*. Urbana and Chicago: University of Illinois Press.

Thorne, K. S. (1988). "Near Zero: New frontiers of Physics". In: ed. by J. D. Fairbank, B. S. Jr Deaver, C. W. F. Everitt, and P. F. Michelson. New York: W H Freeman and Co. Chap. "Gravitomagnetism, Jets in Quasars and the Stanford Gyroscope Experiment", pp. 573 –586.

Torretti, Roberto (2000). "Spacetime Models for the World". In: *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 31.2, pp. 171 –186. ISSN: 1355-2198. DOI: http://dx.doi.org/10.1016/S1355-2198(99)00036-2. URL: http://www.sciencedirect.com/science/article/pii/S1355219899000362.

Totsuka, Y. (1991). "Neutrino Physics". In: ed. by K. Winter. Vol. 14. Cambridge Monographs on Particle Physics, Nuclear Physics and Cosmology. Cambridge: Cambridge University Press. Chap. Supernovae and neutrinos.

Turneaure, J.P., C.W.F. Everitt, B.W. Parkinson, D. Bardas, S. Buchman, D.B. DeBra, H. Dougherty, D. Gill, J. Grammer, G.B. Green, G.M. Gutt, D.-H. Gwo, M. Heifetz, N.J. Kasdin, G.M. Keiser, J.A. Lipa, J.M. Lockhart, J.C. Mester, Barry Muhlfelder, R. Parmley, A.S. Silbergleit, M.T. Sullivan, M.A. Taber, R.A. Van Patten, R. Vassar, S. Wang, Y.M. Xiao, and P. Zhou (2003). "Development of the Gravity Probe B flight mission". In: *Advances in Space Research* 32.7, pp. 1387 –1396. ISSN: 0273-1177. DOI: http://dx.doi.org/10.1016/S0273-1177(03)90351-6. URL: http://www.sciencedirect.com/science/article/pii/S0273117703903516.

Uzan, Jean-Philippe (2003). "The fundamental constants and their variation: observational and theoretical status". In: *Review of Modern Physics* 75, pp. 403–455. DOI: 10.1103/RevModPhys.75.403. URL: http://link.aps.org/doi/10.1103/RevModPhys.75.403.

Van Orman Quine, W. and J. S. Ullian (1970). *The Web of Belief*. Random House.

Van Patten, Richard A., Ray DiEsposti, and John V. Breakwell (1986). "Ultra High Resolution Science Data Extraction For The Gravity Probe-B Gyro And Telescope". In: *Proc. SPIE* 0619, pp. 157–165. DOI: 10.1117/12.966649. URL: http://dx.doi.org/10.1117/12.966649.

Weinberg, Alvin M. (1961). "Impact of Large-Scale Science on the United States". In: *Science* 134.3474, pp. 161–164. DOI: 10.1126/science.134.3473.161.

Will, Clifford M. (1993a). *Theory and Experiment in Gravitational Physics*. Cambridge: Cambridge University Press. ISBN: Revised edition.

– (1993b). *Was Einstein Right? Putting General Relativity to the Test*. New York: Basic Books. ISBN: 9780465090860.

– (2006). "The Confrontation between General Relativity and Experiment". In: *Living Reviews in Relativity* 9.3. DOI: 10.1007/lrr-2006-3. URL: http://www.livingreviews.org/lrr-2006-3.

– (2014). "The Confrontation between General Relativity and Experiment". In: *Living Reviews in Relativity* 17.4. DOI: 10.1007/lrr-2014-4. URL: http://www.livingreviews.org/lrr-2014-4.

– (2015). "Focus issue: Gravity Probe B". In: *Classical and Quantum Gravity* 32.22, p. 220301. URL: http://stacks.iop.org/0264-9381/32/i=22/a=220301.

Woodward, James (1989). "Data and phenomena". In: *Synthese* 79, 393–472.

– (2000). "Data, phenomena, and reliability". In: *Philosophy of Science* 67, S163–S179. DOI: 10.1086/392817.

– (2011). "Data and phenomena: a restatement and defense". In: *Synthese* 182.1, pp. 165–179. DOI: 10.1007/s11229-009-9618-5.

Zahar, Elie (1973). "Why Did Einstein's Programme Supersede Lorentz's? (II)". In: *British Journal for the Philosophy of Science* 24.3, pp. 223–262.