# Roger Penrose Research Papers

Sir Roger Penrose (born 8 August 1931) is an English mathematical physicist, mathematician and philosopher of science. He is Emeritus Rouse Ball Professor of Mathematics at the University of Oxford and an emeritus fellow of Wadham College, Oxford. Penrose has made contributions to the mathematical physics of general relativity and cosmology. He has received several prizes and awards, including the 1988 Wolf Prize for physics, which he shared with Stephen Hawking for the Penrose–Hawking singularity theorems.

| Sir **Roger Penrose** | |
|---|---|
| Penrose in 2011 | |
| **Born** | 8 August 1931 (age 88)<br>Colchester, England |
| **Residence** | England |
| **Nationality** | English |
| **Alma mater** | University College, London<br>St John's College, Cambridge |
| **Known for** | Twistor theory<br>Spin network<br>Abstract index notation<br>Black hole bomb<br>Geometry of spacetime |

| | |
|---|---|
| | Cosmic censorship<br>Weyl curvature hypothesis<br>Penrose inequalities<br>Penrose interpretation of quantum mechanics<br>Moore–Penrose pseudoinverse<br>Newman–Penrose formalism<br>Penrose diagram<br>Penrose inequality<br>Penrose process<br>Penrose tiling<br>Penrose stairs<br>Penrose graphical notation<br>Penrose transform<br>Penrose-Terrell effect<br>pp-wave spacetime<br>Schrödinger–Newton equations<br>Orch-OR/Penrose–Lucas argument<br>FELIX experiment<br>Trapped surface<br>Andromeda paradox<br>Conformal cyclic cosmology |
| **Awards** | Adams Prize 1966<br>Heineman Prize (1971)<br>Fellow of the Royal Society (1972)<br>Eddington Medal (1975)<br>Royal Medal (1985) |

| | |
|---|---|
| | Wolf Prize (1988)<br>Dirac Medal (1989)<br>Albert Einstein Medal (1990)<br>Naylor Prize and Lectureship (1991)<br>James Scott Prize Lectureship (1997–2000)<br>Karl Schwarzschild Medal (2000)<br>De Morgan Medal (2004)<br>Copley Medal (2008)<br>Fonseca Prize (2011) |
| **Scientific career** | |
| **Fields** | Mathematical physics, tessellations |
| **Institutions** | Bedford College, London<br>Princeton University<br>Syracuse University<br>King's College, London<br>Birkbeck, University of London<br>Wadham College, University of Oxford<br>Polish Academy of Sciences |
| **Thesis** | *Tensor Methods in Algebraic Geometry* (1958) |
| **Doctoral advisor** | John A. Todd |
| **Other academic advisors** | W. V. D. Hodge |
| **Doctoral** | Claude LeBrun |

| | |
|---|---|
| **students** | Tristan Needham<br>Richard Jozsa<br>Richard S. Ward<br>Andrew Hodges<br>Asghar Qadir<br>Lane P. Hughston<br>Tim Poston |
| **Influenced** | Michael Atiyah<br>Stuart Hameroff |

arXiv:math/0001097v1 [math.MG] 18 Jan 2000

# Drawing with Complex Numbers

Michael Eastwood[♯] and Roger Penrose

It is not commonly realized that the algebra of complex numbers can be used in an elegant way to represent the images of ordinary 3-dimensional figures, orthographically projected to the plane. We describe these ideas here, both using simple geometry and setting them in a broader context.

Consider orthogonal projection in Euclidean $n$-space onto an $m$-dimensional subspace. We may as well choose coördinates so that this is the standard projection $P : \mathbb{R}^n \to \mathbb{R}^m$ onto the first $m$ variables. Fix a non-degenerate simplex $\Sigma$ in $\mathbb{R}^n$. Two such simplices are said to be similar if one can be obtained from the other by a Euclidean motion together with an overall scaling. This article answers the following question. Given $n+1$ points in $\mathbb{R}^m$, when can these points be obtained as the images under $P$ of the vertices of a simplex similar to $\Sigma$?

When $n = 3$ and $m = 2$, then $P$ is the standard orthographic projection (as often used in engineering drawing) and we are concerned with how to draw a given tetrahedron. We shall show, for example, that four points $\alpha, \beta, \gamma, \delta$ in the plane are the orthographic projections of the vertices of a regular tetrahedron if and only if

$$(\alpha + \beta + \gamma + \delta)^2 = 4(\alpha^2 + \beta^2 + \gamma^2 + \delta^2) \tag{1}$$

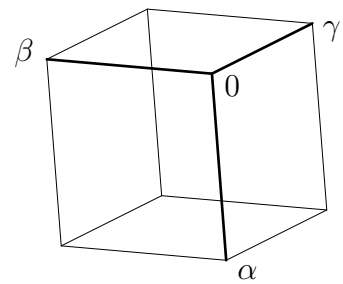
where $\alpha, \beta, \gamma, \delta$ are regarded as complex numbers! Similarly, suppose a cube is orthographically projected and normalised so that a particular vertex is mapped to the origin. If $\alpha, \beta, \gamma$ are the images of the three neighbouring vertices, then

$$\alpha^2 + \beta^2 + \gamma^2 = 0, \tag{2}$$

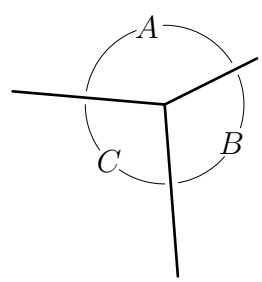
again as a complex equation. Conversely, if this equation is satisfied, then one can find a cube whose orthographic image is given in this way. Since parallel lines are seen as parallel in the drawing, equation (2) allows one to draw the general cube:

[♯]Supported by the Australian Research Council.

In this example, $\alpha = 2 - 26i$
$\beta = -23 + 2i$
$\gamma = 14 + 7i$




The result for a cube is known as Gauss' fundamental theorem of axonometry—see [3, p. 309] where it is stated without proof. In engineering drawing, one usually fixes three principal axes in Euclidean three-space and then an orthographic projection onto a plane transverse to these axes is known as an axonometric projection (see, for example, [8, Chapter 17]). Gauss' theorem may be regarded as determining the degree of foreshortening along the principal axes for a general axonometric projection. The projection corresponding to taking $\alpha, \beta, \gamma$ to be the three cube roots of unity is called isometric projection because the foreshortening is the same for the three principal axes. In an axonometric drawing, it is conventional to take the image axes at mutually obtuse angles:

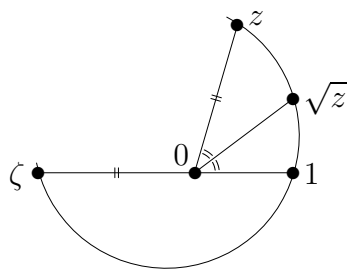



If $|\alpha| = a$, $|\beta| = b$, $|\gamma| = c$, then equation (2) is equivalent to the sine rule for the triangle with sides $\alpha^2$, $\beta^2$, $\gamma^2$, namely

$$\frac{a^2}{\sin 2A} = \frac{b^2}{\sin 2B} = \frac{c^2}{\sin 2C}.$$

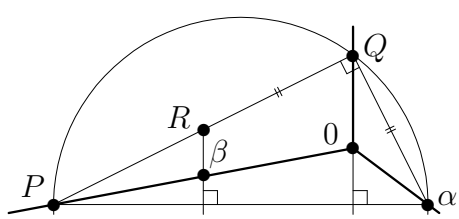
In this form, the fundamental theorem of axonometry is due to Weisbach and was published in Tübingen in 1844 in the Polytechnische Mitteilungen of Volz and Karmasch. Equivalent statements can be found in modern engineering drawing texts (e.g. [7, p. 44]).

Equation (2) may be used to give a ruler and compass construction of the general orthographic image of a cube. If we suppose that the image of a vertex and two of its neighbours are already specified, then (2) determines (up to a two-fold ambiguity) the image of the third neighbour. The construction is straightforward except perhaps for the construction of a complex square root for which we advocate the following as quite efficient:




Firstly $\zeta$ is constructed by marking the real axis at a distance $\|z\|$ from the origin. Then, a circle is constructed passing through the three points $\zeta$, 1, and $z$. Finally, the angle between 1 and $z$ is bisected and $\sqrt{z}$ appears where this bisector meets the circle.

In engineering drawing, it is more usual that the images of the three principal axes are prescribed or chosen by the designer and one needs to determine the relative degree of foreshortening along these axes. There is a ruler and compass construction given by T. Schmid in 1922 (see, for example, [8, §17.17–17.19]):




In this diagram, the three principal axes and $\alpha$ are given. By drawing a perpendicular from $\alpha$ to one of the of the principal axes and marking its intersection with the remaining principal axis, we obtain $P$. The point $Q$ is obtained by drawing a semi-circle as illustrated. The point $R$ is on the resulting line and equidistant with $\alpha$ from $Q$. Finally, $\beta$ is obtained by dropping a perpendicular as shown. It is easy to see that this construction has

the desired effect—in Euclidean three-space rotate the right-angled triangle with hypotenuse $P\alpha$ about this hypotenuse until the point $Q$ lies directly above 0 in which case $R$ will lie directly above $\beta$ and the third vertex will lie somewhere over the line through 0 and $Q$. One may verify the appropriate part of Weisbach's condition

$$\frac{a^2}{\sin 2A} = \frac{b^2}{\sin 2B} \tag{3}$$

by the following calculation. Without loss of generality we may represent all these points by complex numbers normalised so that $Q = 1$. Then it is straightforward to check that

$$R{=}1 + i - i\alpha,\ P{=}\frac{\alpha(\alpha+\overline{\alpha}) + 2(1-\alpha-\overline{\alpha})}{\alpha-\overline{\alpha}},\ \beta{=}\frac{\alpha(\alpha+\overline{\alpha}) + 2(1-\alpha-\overline{\alpha})}{2-\alpha-\overline{\alpha}}i$$

and therefore that

$$\alpha^2 + \beta^2 = 4\frac{(\alpha-1)\overline{(\alpha-1)}(\alpha+\overline{\alpha}-1)}{(\alpha+\overline{\alpha}-2)^2}.$$

That $\alpha^2 + \beta^2$ is real is equivalent to (3).

To prove Gauss' theorem more directly consider three vectors in $\mathbb{R}^3$ as the columns of a $3 \times 3$ matrix. This matrix is orthogonal if and only if the three vectors are orthonormal. It is equivalent to demand that the three rows be orthonormal. However, any two orthonormal vectors in $\mathbb{R}^3$ may be extended to an orthonormal basis. Thus, the condition that three vectors

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} \qquad \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \qquad \begin{pmatrix} x_3 \\ y_3 \end{pmatrix}$$

in $\mathbb{R}^2$ be the images under $P : \mathbb{R}^3 \to \mathbb{R}^2$ of an orthonormal basis of $\mathbb{R}^3$, is that

$$\begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} y_1 & y_2 & y_3 \end{pmatrix}$$

be orthonormal in $\mathbb{R}^3$. Dropping the overall scale, we obtain

$${x_1}^2 + {x_2}^2 + {x_3}^2 = {y_1}^2 + {y_2}^2 + {y_3}^2 \quad \text{and} \quad x_1y_1 + x_2y_2 + x_3y_3 = 0.$$

Writing, $\alpha = x_1 + iy_1$, $\beta = x_2 + y_2$, $\gamma = x_3 + y_3$, these two equations are the real and imaginary parts of (2). To deduce the case of a regular tetrahedron as

described by equation (1) from the case of a cube as described by equation (2), it suffices to note that equation (1) is translation invariant and that a regular tetrahedron may be inscribed in a cube. Thus, we may take $\delta = \alpha + \beta + \gamma$ and observe that (1) and (2) are then equivalent.

It is easy to see that the possible images of a particular tetrahedron $\Sigma$ in $\mathbb{R}^3$ under an arbitrary Euclidean motion followed by the projection $P$ form a 5-dimensional space—the group of Euclidean motions is 6-dimensional but translation orthogonal to the plane leaves the image unaltered. It therefore has codimension 3 in the 8-dimensional space of all tetrahedral images (2 degrees of freedom for each vertex). Allowing similar tetrahedra rather than congruent reduces the codimension to 2. Therefore, two real equations are to be expected. Always, these two real equations combine as a single complex equation such as (1) or (2). At first sight, this is perhaps surprising and even more so when the same phenomenon occurs for $P : \mathbb{R}^n \to \mathbb{R}^2$ for arbitrary $n$.

For $n = 3$, there is a proof of Gauss' theorem which brings in complex numbers at the outset. Consider the space $H$ of Hermitian $2 \times 2$ matrices with zero trace, i.e. matrices of the form

$$X = \begin{pmatrix} w & u + iv \\ u - iv & -w \end{pmatrix} \quad \text{for} \quad \begin{pmatrix} u \\ v \\ w \end{pmatrix} \in \mathbb{R}^3.$$

We may identify $H$ with $\mathbb{R}^3$ and, in so doing, $-\det X$ becomes the square of the Euclidean length. The group $G$ of invertible $2 \times 2$ complex matrices of the form

$$\Lambda = \begin{pmatrix} a & -b \\ \overline{b} & \overline{a} \end{pmatrix}$$

acts linearly on $H$ by $X \mapsto \Lambda X \overline{\Lambda}^t$. Moreover,

$$\det(\Lambda X \overline{\Lambda}^t) = (|a|^2 + |b|^2)^2 \det X$$

so $G$ acts by similarities. It is easy to check that all similarities may be obtained in this way. (This trick is essentially as used in Hamilton's theory of quaternions and is well known to physicists—in modern parlance it is equivalent to the isomorphism of Lie groups $\mathrm{Spin}(3) \cong \mathrm{SU}(2)$.) Therefore, an arbitrary orthographicimage of a cube may be obtained by acting with $\Lambda$ on the standard basis

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

and then picking out the top right hand entries. We obtain

$$\Lambda \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \overline{\Lambda}^t = \begin{pmatrix} * & a^2 - b^2 \\ * & * \end{pmatrix} \longmapsto a^2 - b^2 = \alpha$$

$$\Lambda \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix} \overline{\Lambda}^t = \begin{pmatrix} * & i(a^2 + b^2) \\ * & * \end{pmatrix} \longmapsto i(a^2 + b^2) = \beta$$

$$\Lambda \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \overline{\Lambda}^t = \begin{pmatrix} * & 2ab \\ * & * \end{pmatrix} \longmapsto 2ab = \gamma$$

and therefore $\alpha^2 + \beta^2 + \gamma^2 = 0$, as required. Conversely, this is exactly the condition that $\alpha, \beta, \gamma$ may be written in this form. (Compare the half angle formulae—if $s^2 + c^2 = 1$, then $s = 2t/(1 + t^2)$ and $c = (1 - t^2)/(1 + t^2)$ for some $t$.) That Gauss [3, p. 309] makes the same observation concerning the form of $\alpha, \beta, \gamma$ suggests that perhaps he also had this reasoning in mind.

The proof of Gauss' theorem using orthogonal matrices clearly extends to $P : \mathbb{R}^n \to \mathbb{R}^2 = \mathbb{C}$ for arbitrary $n$. To state it, the following terminology concerning the standard projection $P : \mathbb{R}^n \to \mathbb{R}^m$ is useful. We shall say that $v_1, v_2, \ldots, v_n \in \mathbb{R}^m$ are <u>normalised</u> <u>eutactic</u> if and only if there is an orthonormal basis $u_1, u_2, \ldots, u_n$ of $\mathbb{R}^n$ with $v_j = Pu_j$ for $j = 1, 2, \ldots, n$. We shall say that $v_1, v_2, \ldots, v_n \in \mathbb{R}^m$ are <u>eutactic</u> if and only if $\mu v_1, \mu v_2, \ldots, \mu v_n$ are normalised eutactic for some $\mu \neq 0$.

**Theorem** *The points $z_1, z_2, \ldots, z_n \in \mathbb{C} = \mathbb{R}^2$ are eutactic if and only if*

$$z_1{}^2 + z_2{}^2 + \cdots + z_n{}^2 = 0$$

*and not all $z_j$ are zero.*

There is a proof for $n = 4$ based on the isomorphism

$$\mathrm{Spin}(4) \cong \mathrm{SU}(2) \times \mathrm{SU}(2)$$

and, indeed, this is how we came across the theorem in the first place. However, a more direct route to complex numbers and one which applies in all dimensions is based on the observation that $\mathrm{Gr}_2^+(\mathbb{R}^2)$, the Grassmannian of oriented two-planes in $\mathbb{R}^n$, is naturally a <u>complex</u> manifold. When $n = 3$,

this Grassmannian is just the two-sphere and has a complex structure as the Riemann sphere. In general, consider the mapping

$$\mathbb{CP}_{n-1} \setminus \mathbb{RP}_{n-1} \xrightarrow{\pi} \mathrm{Gr}_2^+(\mathbb{R}^n)$$

induced by $\mathbb{C}^n \ni z \mapsto iz \wedge \overline{z}$. In other words, a complex vector $z = x+iy \in \mathbb{C}^n$ is mapped to the two-dimensional oriented subspace of $\mathbb{R}^n$ spanned by $x$ and $y$, the real and imaginary parts of $z$. Let $\langle\ ,\ \rangle$ denote the standard inner product on $\mathbb{R}^n$ extended to $\mathbb{C}^n$ as a complex bilinear form. Then, $\langle z, z\rangle = 0$ imposes two real equations

$$\|x\|^2 = \|y\|^2 \quad \text{and} \quad \langle x, y\rangle = 0$$

on the real and imaginary parts. In other words, $x, y$ is proportional to an orthonormal basis for $\mathrm{span}\{x, y\}$. Hence, if $z$ and $w$ satisfy $\langle z, z\rangle = 0 = \langle w, w\rangle$ and define the same oriented two-plane, then $w = \lambda z$ for some $\lambda \in \mathbb{C} \setminus \{0\}$. The non-singular complex quadric

$$K = \{[z] \in \mathbb{CP}_{n-1} \text{ s.t. } \langle z, z\rangle = 0\}$$

avoids $\mathbb{RP}_{n-1} \subset \mathbb{CP}_{n-1}$ and we have shown that $\pi|_K$ is injective. It is clearly surjective. The isomorphism

$$\pi : K \xrightarrow{\cong} \mathrm{Gr}_2^+(\mathbb{R}^n)$$

respects the natural action of $\mathrm{SO}(n)$ on $K$ and $\mathrm{Gr}_2^+(\mathbb{R}^n)$. The generalised Gauss theorem follows immediately since, rather than asking about the image of a general orthonormal basis under the standard projection $P : \mathbb{R}^n \to \mathbb{R}^2$, we may, equivalently, ask about the image of the standard basis $e_1, e_2, \ldots, e_n$ under a general orthogonal projection onto an oriented two-plane $\Pi \subset \mathbb{R}^n$. Any such $\Pi$ is naturally complex, the action of $i$ being given by rotation by $90°$ in the positive sense. If $\Pi$ is represented by $[z_1, z_2, \ldots, z_n] \in K$ as above and we use $x, y \in \Pi$ to identify $\Pi$ with $\mathbb{C}$, then $e_j \mapsto z_j$ and

$${z_1}^2 + {z_2}^2 + \cdots + {z_n}^2 = \langle z, z\rangle = 0,$$

as required. Conversely, a solution of this complex equation determines an appropriate plane $\Pi$.

For the case of a general tetrahedron or simplex and for general $m$ and $n$, it is more convenient to start with Hadwiger's theorem [4] or [2, page 251] as follows. The proof is obtained by extending our orthogonal matrix proof of Gauss' theorem.

**Theorem (Hadwiger)** *Assemble $v_1, v_2, \ldots, , v_n \in \mathbb{R}^m$ as the columns of an $m \times n$ matrix $V$. These vectors are normalised eutactic if and only if $VV^t = 1$ (the $m \times m$ identity matrix).*

**Proof** If $v_1, v_2, \ldots, v_n$ are normalised eutactic, then assembling a corresponding orthonormal basis of $\mathbb{R}^n$ as the columns of an $n \times n$ matrix, we have $V = PU$ and $U^tU = 1$ (the $n \times n$ identity matrix). Therefore, $UU^t = 1$ and

$$VV^t = PUU^tP^t = PP^t = 1,$$

as required. Conversely, if $VV^t = 1$, then the columns of $V^t$ may be completed to an orthonormal basis of $\mathbb{R}^n$, i.e. $V^t = U^tP^t$ for $UU^t = 1$. Now, $U^tU = 1$ and $V = PU$, as required. □

The case of a general simplex is obtained essentially by a change of basis as follows. Suppose $a_1, a_2, \ldots, a_n, a_{n+1}$ are the vertices of a non-degenerate simplex $\Sigma$ in $\mathbb{R}^n$ whose centre of mass is at the origin. In other words, the $n \times (n+1)$ matrix $A$ has rank $n$ and $Ae = 0$ where $e$ is the column vector all of whose $n+1$ entries are 1. Form the $(n+1) \times (n+1)$ symmetric matrix

$$Q = A^t(AA^t)^{-2}A,$$

noting that $\operatorname{rank} A = n$ implies the <u>moment</u> <u>matrix</u> $AA^t$ is invertible.

**Theorem** *Given $b_1, b_2, \ldots, b_n, b_{n+1} \in \mathbb{R}^m$ assembled as the columns of an $m \times (n+1)$ matrix $B$, these vectors are the images under orthogonal projection $P : \mathbb{R}^n \to \mathbb{R}^m$ of the vertices of a simplex congruent to $\Sigma$ if and only if*

$$BQ^tB = 1. \tag{4}$$

**Proof** The vertices of a simplex congruent to $\Sigma$ are the columns of a matrix $UA + ae^t$ for some orthogonal matrix $U$ and translation vector $a \in \mathbb{R}^n$. Also, note that $Qe = 0$. Thus, if $B = P(UA + ae^t)$, then

$$\begin{aligned} BQB^t &= PUAQA^tU^tP^t \\ &= PUAA^t(AA^t)^{-2}AA^tU^tP^t \\ &= PUU^tP^t = PP^t = 1, \end{aligned}$$

as required. Conversely, $Qe = 0$ implies that (4) is translation invariant. So, without loss of generality, we may suppose that $b_1 + b_2 + \cdots + b_n + b_{n+1} = 0$, that is to say, $Be = 0$. Writing out (4) in full gives

$$BA^t(AA^t)^{-1}(BA^t(AA^t)^{-1})^t = 1$$

so, by Hadwiger's theorem, there is an orthogonal matrix $U$ so that

$$BA^t(AA^t)^{-1} = PU.$$

Thus,

$$BA^t(AA^t)^{-1}A = PUA \quad \text{and} \quad Be = 0.$$

Certainly, $B = PUA$ is a solution of these equations but it is the only solution since $A^t(AA^t)^{-1}A$ has rank $n$ and $e$ is not in the range of this linear transformation. □

**Corollary (case m = 2)** *Points $z_1, z_2, \ldots, z_n, z_{n+1} \in \mathbb{C}$ are the images under orthogonal projection of the vertices of a simplex similar to $\Sigma$ if and only if*

$$z^t Q z = 0$$

*where $z$ is the column vector with components $z_1, z_2, \ldots, z_n, z_{n+1}$.*

It is, of course, possible to compute $Q$ explicitly for any given example. If the simplex $\Sigma$ has some degree of symmetry, however, we can often circumvent such computation. Consider, for example, the case of a regular simplex. From the corollary above, we know that the image of such a simplex in the plane is characterised by a complex homogeneous quadratic polynomial. The symmetries of the regular simplex ensure that this polynomial must be invariant under $\mathcal{S}_{n+1}$, the symmetric group on $n+1$ letters. Hence, it must be expressible in terms of the elementary symmetric polynomials. Equivalently, it must be a linear combination of

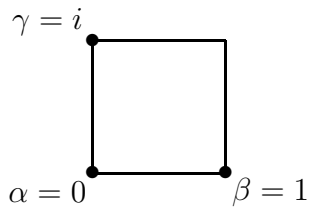
$$(z_1 + z_2 + \cdots + z_n + z_{n+1})^2 \quad \text{and} \quad {z_1}^2 + {z_2}^2 + \cdots + {z_n}^2 + {z_{n+1}}^2.$$

Up to scale, there is only one such combination which is translation invariant, namely

$$(z_1 + z_2 + \cdots + z_n + z_{n+1})^2 - (n+1)({z_1}^2 + {z_2}^2 + \cdots + {z_n}^2 + {z_{n+1}}^2). \quad (5)$$

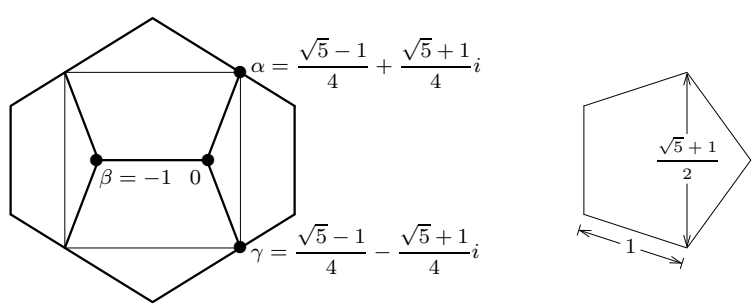
It follows that the vanishing of this polynomial is an equation which characterises the possible images of a regular simplex under orthogonal projection into the plane. The special case $n = 2$ characterises the equilateral triangles in the plane [1, Problem 15 on page 79].

Equation (2) characterising the orthographic images of a cube, may be deduced by similar symmetry considerations. If a particular vertex is mapped to the origin and its neighbours are mapped to $\alpha, \beta, \gamma$ then, since each of these neighbouring vertices is on an equal footing, the polynomial in question must be a linear combination of $(\alpha + \beta + \gamma)^2$ and $\alpha^2 + \beta^2 + \gamma^2$. To find out which linear combination we need only consider a particular projection, for example:




In this example, $(\alpha + \beta + \gamma)^2 = 2i$ and $\alpha^2 + \beta^2 + \gamma^2 = 0$. Up to scale, therefore, (2) is the correct equation.

The case of a regular dodecahedron is similar. Using the fact that a cube may be inscribed in such a dodecahedron [5], we may deduce a particular projection:




with $(\alpha + \beta + \gamma)^2 = (7 - 3\sqrt{5})/2$ and $\alpha^2 + \beta^2 + \gamma^2 = (2 - \sqrt{5})/2$. In this particular case,

$$(\alpha + \beta + \gamma)^2 + (\sqrt{5} - 1)(\alpha^2 + \beta^2 + \gamma^2) = 0.$$

Therefore, this is the correct equation in the general case. It may be used as the basis of a ruler and compass construction of the general orthographic projection of a regular dodecahedron.

It is interesting to note that if all the vertices of a Platonic solid are orthographically projected to $z_1, z_2, \ldots, z_N \in \mathbb{C}$, then

$$(z_1 + z_2 + \cdots + z_N)^2 = N({z_1}^2 + {z_2}^2 + \cdots + {z_N}^2) \tag{6}$$

(compare (5)). For a tetrahedron, this is just equation (1). To verify (6) for the other Platonic solids, firstly note that it is translation invariant. Therefore, it suffices to impose $z_1 + z_2 + \cdots + z_N = 0$ and show that ${z_1}^2 + {z_2}^2 + \cdots + {z_N}^2 = 0$. The case of a cube now follows immediately since its vertices may be grouped as two regular tetrahedra. The dodecahedral case may be dealt with by grouping its vertices into five regular tetrahedra. The regular octahedron is amenable to a similar trick but not the icosahedron. Rather than resorting to direct computation, a uniform proof may be given as follows. As before, assemble the vertices of the given solid $\Sigma$ as the columns of a matrix $A$, now of size $3 \times N$, and consider the moment matrix $M \equiv AA^t$. Observe that

$$\begin{pmatrix} 1 & i & 0 \end{pmatrix} M \begin{pmatrix} 1 \\ i \\ 0 \end{pmatrix} = {z_1}^2 + {z_2}^2 + \cdots + {z_N}^2.$$

The moment matrix is positive definite and symmetric. In other words, it defines a metric on $\mathbb{R}^3$, manifestly invariant under the symmetries of $\Sigma$. If $\Sigma$ is regular—or, more generally, enjoys the symmetries of a regular solid (e.g. a cuboctahedron or rhombicosidodecahedron)—then its symmetry group acts irreducibly on $\mathbb{R}^3$. Thus, $M$ must be proportional to the identity matrix and the result follows. For a general solid $\Sigma$, the two complex numbers

$$\pm\sqrt{{z_1}^2 + {z_2}^2 + \cdots + {z_N}^2}$$

are the foci of the ellipse

$$\begin{pmatrix} x & y \end{pmatrix} R \begin{pmatrix} x \\ y \end{pmatrix} = 1$$

where $R$ is the inverse of the quadratic form obtained by restricting M to the plane of projection.

This reasoning also works in higher dimensions where it shows (as conjectured to us by H.S.M. Coxeter) that the orthogonally projected images in the plane of the $N$ vertices of any regular polytope, real or complex, will satisfy equation (6). Of course, this excludes regular polygons (whose symmetry groups act reducibly except in dimension two) orthographic images of which will satisfy (6) if and only if the image is itself regular. For polyhedra other than simplices, a quadratic equation such as (6) is no longer sufficient to characterise the orthogonal image up to scale. In general, there will also be some linear relations. For a non-degenerate $N$-gon there will be $N-n-1$ such relations. The simplest example is a square in $\mathbb{R}^2$ which is characterised by the complex equations

$$(\alpha+\beta+\gamma+\delta)^2 = 4(\alpha^2+\beta^2+\gamma^2+\delta^2) \quad \text{and} \quad \alpha+\gamma=\beta+\delta.$$

It is interesting to investigate further the relationship between a non-degenerate simplex $\Sigma$ in $\mathbb{R}^n$ and its quadratic form $Q = A^t(AA^t)^{-2}A$. Recall that $A$ is the $n\times(n+1)$ matrix whose columns are the vertices of $\Sigma$. There are several other formulae for or characterisations of $Q$. Let $S$ denote the $(n+1)\times(n+1)$ symmetric matrix

$$1 - \frac{1}{n+1}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}.$$

It is the matrix of orthogonal projection in $\mathbb{R}^{n+1}$ in the direction of the vector $e$. We maintain that $Q$ is characterised by the equations

$$QA^tA = S \quad \text{and} \quad Qe = 0.$$

Certainly, if these equations hold, then they are enough to determine $Q$ because the moment matrix $M \equiv A^tA$ has rank $n$ and $e$ is not in its range. The second equation is evident and the first equation with $Q$ replaced by $A^t(AA^t)^{-2}A$ and simplified reads

$$A^t(AA^t)^{-1}A = S.$$

To see that this holds it suffices to observe that it is clearly true after post-multiplication by $A^t$ or $e$. We may equally well characterise $Q$ by means of the equations

$$A^tAQ = S \quad \text{and} \quad Qe = 0.$$

These equations relate $M$ and $Q$ geometrically—both matrices annihilate $e$ whilst on the hyperplane orthogonal to $e$ they are mutually inverse. This implies that $M$ and $Q$ are generalised inverses [6] of each other. Thus,

$$Q = M^\dagger = (A^t A)^\dagger = A^\dagger A^{\dagger t}$$

where $A^\dagger$ is the generalised inverse of $A$. In this case, $A^\dagger = A^t(AA^t)^{-1}$. This also shows how to compute $Q$ more directly in certain cases. The moment matrix $M$ has direct geometric interpretation as the various inner products of the vectors $a_1, a_2, \ldots, a_n, a_{n+1}$. In the case of a regular simplex, for example, we know that $\|a_i\|^2$ is independent of $i$, that $\|a_i - a_j\|^2$ is independent of $i \neq j$, and that $a_1 + a_2 + \ldots + a_n + a_{n+1} = 0$. We may deduce that, with a suitable overall scale, $M = S$. Since $S^\dagger = S$, it follows that $Q = S$. This is a direct derivation of (5).

It is clear geometrically that $M$ or, equivalently, $Q$ determines $\Sigma$ up to congruency. Alternatively, one can argue algebraically—it is easy to check that if $A^t A = B^t B$, then $U = AA^t(BA^t)^{-1}$ is orthogonal and $A = UB$. Therefore, the possible quadratic forms $Q$ which can arise give a natural parametrisation of the non-degenerate simplices up to congruency. Choosing a basepoint $\Sigma_0$ with corresponding matrix $A_0$, and mapping $X \in \mathrm{GL}(n, \mathbb{R})$ to $X^{-1}A_0$ identifies the space of non-degenerate simplices up to congruency with the homogeneous space $\mathrm{GL}(n, \mathbb{R})/\mathrm{O}(n)$. This homogeneous space may also be identified with the space of positive definite $n \times n$ quadratic forms by sending $X \in \mathrm{GL}(n, \mathbb{R})$ to $XX^t$. The $(n+1) \times (n+1)$ quadratic form $Q$ corresponding to $X^{-1}A_0$ is given by $A_0^\dagger X X^t A_0^{\dagger t}$. It follows that the general $Q$ which can arise is characterised by the following two conditions:

- $Qe = 0$ and only multiples of $e$ are in the kernel of $Q$.
- All other eigenvalues of $Q$ are positive.

It is also possible to repeat this analysis in pseudo-Euclidean spaces. The only difference is that the condition that the non-zero eigenvalues of $Q$ be positive is replaced by a condition on sign precisely reflecting the original signature of the inner product.

Finally we should mention some possible applications. There is much current interest in computer vision. In particular, there is the problem of recognising a wire-frame object from its orthographic image. The results we

have described can be used as test on such an image, for example to see whether a given image could be that of a cube or to keep track of a moving shape. It is clear that such tests could be implemented quite efficiently. Another possibility is in the manipulation of CADD[♯] data. Rather than storing an image as an array of vectors in $\mathbb{R}^3$, it may be sometimes be more efficient to store certain tetrahedra within such an image by means of the corresponding quadratic form. For orthographic imaging this may be preferable.

We would like to thank H.S.M. Coxeter for drawing our attention to Hadwiger's article, R. Michaels and J. Cofman for pointing out Gauss' and Weisbach's work, and E.J. Pitman for many useful conversations.

## References

[1] S. Barnard and J.M. Child, *Higher Algebra*, MacMillan 1936.

[2] H.S.M. Coxeter, *Regular Polytopes*, Methuen 1948.

[3] C.F. Gauss, *Werke, Zweiter Band*, Königlichen Gesellschaft der Wissenschaften, Göttingen 1876.

[4] H. Hadwiger, *Über ausgezeichnete Vectorsterne und reguläre Polytope*, Comment. Math. Helv. **13** (1940), 90–108.

[5] D. Hilbert and S. Cohn-Vossen, *Geometry and the Imagination*, Chelesa 1952, 1983, 1990.

[6] R. Penrose, *A generalised inverse for matrices*, Proc. Camb. Phil. Soc. **51** (1955), 406–413.

[7] R.N. Roth and I.A. van Haeringen, *The Australian Engineering Drawing Handbook, Part One*, The Institute of Engineers, Australia 1988.

[8] R.P. Hoelscher and C.H. Springer, *Engineering Drawing and Geometry, Second Edition*, Wiley 1961.

Department of Pure Mathematics
University of Adelaide
South AUSTRALIA 5005

Mathematical Institute
24-29 Saint Giles'
Oxford OX1 3LB
ENGLAND

[♯]Computer Aided Drafting and Design.

arXiv:1205.5823v1 [cs.GL] 25 May 2012

# Foreword to **A Computable Universe** Understanding Computation & Exploring Nature As Computation*

Roger Penrose
Mathematical Institute, University of Oxford, UK

I am most honoured to have the privilege to present the Foreword to this fascinating and wonderfully varied collection of contributions, concerning the nature of computation and of its deep connection with the operation of those basic laws, known or yet unknown, governing the universe in which we live. Fundamentally deep questions are indeed being grappled with here, and the fact that we find so many different viewpoints is something to be expected, since, in truth, we know little about the foundational nature and origins of these basic laws, despite the immense precision that we so often find revealed in them. Accordingly, it is not surprising that within the viewpoints expressed here is some unabashed speculation, occasionally bordering on just partially justified guesswork, while elsewhere we find a good deal of precise reasoning, some in the form of rigorous mathematical theorems. Both of these are as should be, for without some inspired guesswork we cannot have new ideas as to where look in order to make genuinely new progress, and without precise mathematical reasoning, no less than in precise observation, we cannot know when we are right—or, more usually, when we are wrong.

The year of the publication of this book, 2012, is particularly apposite, in being the centenary year of Alan Turing, whose theoretical analysis of

*Published in H. Zenil (ed), *A Computable Universe: Understanding Computation & Exploring Nature As Computation*, World Scientific, 2012. Footnotes to names are pointers to the chapters in the volume for which this foreword was written.

the notion of "computing machine", together with his wartime work in deciphering Nazi codes, has had a huge impact on the enormous development of electronic computers, and on the consequent influence that these devices have had on our lives and on the way that we think about ourselves. This impact is particularly evident with the application of computer technology to the implications of known physical laws, whether they be at the basic foundational level, or at a larger level such as with fluid mechanics or thermodynamics where averages over huge numbers of elementary constituent particles again lead to comparatively simple dynamical equations. I should here remark that from time to time it has even been suggested that, in some sense, the "laws" that we appear to find in the way that the world works are *all* of this statistical character, and that, at root, there are "no" basic underlying physical laws (e.g. Wheeler's "law without law" [43], Sakharov's ideas of "induced gravity" [31], etc., and we find this general type of view expressed also in this volume also[1]). However, I find it hard to see that such a viewpoint can have much chance of yielding anything like the enormously precise non-statistical dynamics [32] and great mathematical sophistication that we find in so much of 20th century physics. This point aside, we find that in reasonably favourable circumstances, computer simulations can lead to hugely impressive imitations of reality, and the resulting visual representations may be almost indistinguishable from the real thing, a fact that is frequently made use of in realistic special effects in films, as much as in serious scientific presentations. When we need precision in particular implications of such equations, we may run into the difficult issues presented by chaotic behaviour, whereby the dependence on initial conditions becomes exponentially sensitive. In such cases there is an effective randomness in the evolved behaviour. Nevertheless, the computational simulations will still lead to outcomes that would be physically allowable, and in this sense provide results consistent with the behaviour of reality.

Computational simulations can have great importance in many areas other than physics, such as with the spread of epidemics, or with economics (where the mathematical ideas of game theory can play an important role),[2] but I shall here be concerned with *physical* systems, specifically. The impressiveness of computational simulations is often most evident when it is simply 17th century Newtonian mechanics that is involved, in its enormously varied

[1] see Calude.
[2] Velupillai.

different manifestations. The implications of Newtonian dynamical laws can be extensively computed in the modelling of physical systems, even where there may be huge numbers of constituent particles, such as atoms in a simplified gas, or particle-like ingredients, such as stars in globular clusters or even in entire galaxies. It may be remarked that computational simulations are normally done in a time sense where the future behaviour is deduced from an input which is taken to be in the past. In principle, one could also perform calculations in the reverse "teleological" direction, because of the time-reversibility of the basic Newtonian laws.[3] However, because of the second law of thermodynamics, whereby the entropy (or "randomness") of a physical system increases with time in the natural world, such reverse-time calculations tend to be untrustworthy.

When Newtonian laws are supplemented by the Maxwell-Lorentz equations, governing the behaviour of electromagnetic fields and their interactions with charged material particles, then the scope of physical processes that can be accurately simulated by computational procedures is greatly increased, such as with phenomena involving the behaviour of visible light, or with devices concerned with microwaves or radio propagation, or in modelling the vast galactic plasma clouds involving the mixed flows of electrons and protons in space, which can indeed be computationally simulated with considerable confidence.

This latter kind of simulation requires that those physical equations be used, that correctly come from the requirements of special relativity, where Einstein's viewpoint concerning the relativity of motion and of the passage of time are incorporated. Einstein's special relativity encompassed, encapsulated, and superseded the earlier ideas of FitzGerald, Lorentz, Poincaré and others, but even Einstein's own viewpoint needed to be reformulated and made more satisfactory by the radical change of perspective introduced by Minkowski, who showed how the ideas of special relativity come together in the natural geometrical framework of 4-dimensional space-time. When it comes to Einstein's *general* relativity, in which Minkowski's 4-geometry is fundamentally modified to become *curved*, in order that gravitational phenomena can be incorporated, we find that simulations of gravitational systems can be made to even greater precision than was possible with Newtonian theory. The precision of planetary motions in our Solar System is now at such a level that Newton's 17th century theory is no longer sufficient, and

[3] Beavers.

Einstein's 20th century theory is needed. This is true even for the operation of the global positioning systems that are now in common use, which would be useless but for the corrections to Newtonian theory that general relativity provides. Indeed, perhaps the most accurately confirmed theoretical simulations ever performed, namely the tracking of double neutron-star motions, where not only the standard general-relativistic corrections (perihelion advance, rotational frame-dragging effects, etc.) to Newtonian orbital motion need to be taken into account, but also the energy-removing effects of gravitational waves (ripples of space-time curvature) emanating from the system can be theoretically calculated, and are found to agree with the observed motions to an unprecedented precision.

The other major revolution in basic physical theory that the 20th century brought was, of course, *quantum mechanics*—which needs to be considered in conjunction with its generalization to quantum *field* theory, this being required when the effects of special relativity have to be taken into account together with quantum principles. It is clear from many of the articles in this volume, that quantum theory is (rightly) considered to be of fundamental importance, when it comes to the investigation of the basic underlying operations of the physical universe and their relation to computation. There are many reasons for this, an obvious one being that quantum processes are undoubtedly fundamental to the behaviour of the tiniest-scale ingredients of our universe, and also to many features of the collective behaviour of many-particle systems, these having a characteristically quantum-mechanical nature such as quantum entanglement, superconductivity, Bose-Einstein condensation, etc. However, there is another basic feature of quantum mechanics that may be counted as a reason for regarding this scheme of things as being more friendly to the notion of computation than was classical mechanics, namely that there is a basic *discreteness* that quantum mechanics introduces into physical theory. It seems that in the early days of the theory, much was made of this discreteness, with its implied hope of a "granular" nature underlying the operation of the physical world. A hope had been expressed[30, 32] that somehow the domination of physical theory by the ideas of *continuity* and *differentiability*—which go hand-in-hand with the pervasive use of the real-number system—might have at last been broken, via the introduction of quantum mechanics. Accordingly, it was hoped that the ideas of discreteness and combinatorics might soon be seen to become the dominant driving force underlying the operation of our universe, rather than the continuity and differentiability that classical physics had depended upon for so many

centuries. A discrete universe is indeed much more in harmony with current ideas of computation than is a continuous one, and many of the articles in this volume[4] argue powerfully from this perspective, and particularly in the context of *cellular automata*[5].

The very notion of "computability" that arose from the early 20th century work of various logicians Gödel, Church, Kleene and many others, harking back even to the 19th century ideas of Charles Babbage and Ada Lovelace,[6] and which was greatly clarified by Turing's notion of a computing machine, and by Post's closely related ideas, indeed depend on a fundamental *discreteness* of the basic ingredients. The various very different-looking proposals for a notion of effective computability that these early 20th century logicians introduced all turned out to be *equivalent* to one another, a fact that is central to our current viewpoint concerning computation, and which provides us with the *Church–Turing thesis*, namely that this precise theoretical notion of "computability" does indeed encapsulate the idea of what we intuitively mean by an idealized "mechanical procedure". We find this issue discussed at some depth by numerous authors in this volume[7]. For my own part, I am happy to accept the Church–Turing thesis, in this *original* sense of this phrase, namely that the mathematical notion of computability—as defined by what can be achieved by Church's $\lambda$-calculus, or equivalently by a Turing machine—is indeed the appropriate ideal *mathematical* notion that we require for our considerations of computability. Whether or not the universe in which we live operates in accordance with such a notion of computation is then an issue that we may speculate about, or reason about in one way or another (see, for example, Refs. [45, 20]).

Nevertheless, I can appreciate that there are other viewpoints on this, and that some would prefer to define "computation" in terms of what a physical object can (in principle?) achieve[8]. To me, however, this begs the question, and this same question certainly *remains*, whichever may be our preference concerning the use of the term "computation". If we prefer to use this "physical" definition, then all physical systems "compute" *by definition*, and in that case we would simply need a different word for the (original Church-Turing) mathematical concept of computation, so that the

[4]Bolognesi, Chaitin, Wolfram, Fredkin, and Zenil.
[5]Martínez, Margenstern, Sutner, Wiedermann and Zuse.
[6]DeMol, Sieg, Sutner, Swade and Zuse.
[7]DeMol, Sieg, Dershowitz, Sutner, Bauer and Cooper.
[8]Deutsch, Teuscher, Bauer and Cooper.

profound question raised, concerning the perhaps computable nature of the laws governing the operation of the universe can be studied, and indeed questioned. Accordingly, I shall here use the term "computation" in this *mathematical* sense, and I address this question of the computational nature of physical laws in a serious way later.

Returning, now, to the issue of the discreteness that came through the introduction of standard quantum mechanics, we find that the theory, as we understand it today, has *not* developed in this fundamentally discrete direction that would have fitted in so well with our ideas of computation. The discreteness that Max Planck revealed, in 1900, in his analysis of black-body radiation (although not initially stated in this way) was in effect a discreteness of *phase space*—that high-dimensional mathematical space where each spatial degree of freedom, in a many-particle system, is accompanied by a corresponding momentum degree of freedom. This is not a discreteness that could apply directly to our seemingly continuous perceptions of space and time. Nonetheless, various contributors to this volume[9] have ventured in that more radical direction, arguing that some kind of discreteness might be revealed when we try to examine spatial separations of around the *Planck length* $l_P$ (approximately $10^{-35}$m) and temporal separations of around the Planck time (approximately $10^{-43}$s). These separations are absurdly tiny, smaller by some 20 orders of magnitude from scales of distance and time that are relevant to the processes of standard particle physics. Since these Planck scales are enormously far below anything that modern particle accelerators have been able to explore, it can be reasonably argued that a granularity in the very structure of space-time occurring at the absurdly tiny Planck scales would not have been noticed in current experiments. In addition to this, it has long been argued by some theoreticians, most notably by the distinguished and highly insightful American physicist John A. Wheeler, [42] that our understanding of how a *quantum-gravity* theory ought to operate (according to which the principles of quantum mechanics are imposed upon Einstein's general theory of relativity) tells us that we must indeed expect that at the Planck scales of space-time, something radically new ought to appear, where the smooth space-time picture that we adopt in classical physics would have to be abandoned and something quite different should emerge at this level. Wheeler's argument—based on principles coming from conventional ideas of how Heisenberg's uncertainty principle when applied

[9] Bolognesi and Lloyd.

to quantum fields—involves us in having to envisage wild "quantum fluctuations" that would occur at the Planck scale, providing us with a picture of a seething mess of topological fluctuations. While this picture is not at all similar to that of a discrete granular space-time, it is at least supportive of the idea that something very different from a classically smooth manifold ought to be relevant to Planck-scale physics, and it *might* turn out that a discrete picture is really the correct one. This is a matter that I shall need to return to later in this Foreword.

When it comes to the simulation of conventional quantum systems (not involving anything of the nature of Planck-scale physics) then, as was the case with classical systems, we find that we need to consider the smooth solutions of a (partial) differential equation—in this case the *Schrödinger equation*. Thus, just as with classical dynamics, we cannot directly apply the Church–Turing notion of computability to the evolution of a quantum system, and it seems that we are driven to look for simulations that are mere *approximations* to the exact continuous evolution of Schrödinger's wave function. Turing himself was careful to address this kind of issue, [39] whether it be in the classical or quantum context, and he argued, in effect, that discrete approximations when they are not good enough for some particular purpose can always be improved upon while still remaining discrete. It is indeed one of the key advantages of digital as opposed to analogue representations, that an exponential increase in the accuracy of a digital simulation can be achieved simply by incorporating additional digits. Of course, the simulation could take much longer to run when more digits are included in the approximation, but the issue here is what can *in principle* be achieved by a digital simulation rather than what is practical. In theory, so the argument goes, the discrete approximations can always be increased in accuracy, so that the computational simulations of physical dynamical process can be as precise as would be desired.

Personally, I am not fully convinced by this type of argument, particularly when chaotic systems are being simulated. If we are merely asking for our simulations to represent *plausible* outcomes, consistent with all the relevant physical equations, for the behaviour of some physical system under consideration, then chaotic behaviour may well not be a problem, since we would merely be interested in our simulation being realistic, not that it produces the actual outcome that will in fact come about. On the other hand, if—as in weather prediction—it is indeed required that our simulation emphis to provide the actual outcome of the behaviour of some specific system occurring

in the world that we actually inhabit, then this is another matter altogether, and approximations may not be sufficient, so that chaotic behaviour becomes a genuinely problematic issue.[10]

It may be noted, however, that the Schrödinger equation, being *linear*, does not, strictly speaking, have chaotic solutions. Nevertheless, there is a notion known as "quantum chaos", which normally refers to quantum systems that are the quantizations of chaotic classical systems. Here the issue of "quantum chaos" is a subtle one, and is all tied up with the question of what we normally wish to use the Schrödinger equation *for*, which has to do with the fraught issue of *quantum measurement*. What we find in practice, in a general way—and I shall need to return to this issue later—is that the evolution of the Schrödinger equation does not provide us with the unique outcome that we find to have occurred in the actual world, but with a superposition of possible alternative outcomes, with a probability value assigned to each. The situation is, in effect, no better than with chaotic systems, and again our computational simulations cannot be used to predict the actual dynamical outcome of a particular physical system. As with chaotic systems, all that our simulations give us will be alternative outcomes that are plausible ones—with probability values attached—and will not normally give us a clear prediction of the future behaviour of a particular physical system. In fact, the quantum situation is in a sense "worse" than with classical chaotic systems, since here the lack of predictiveness does not result from limitations on the accuracy of the computational simulations that can be carried out, but we find that even a completely precise simulation of the required solution of the Schrödinger equation would not enable us to predict with confidence what the *actual* outcome would be. The unique history that emerges, in the universe we actually experience, is but one member of the *superposition* that the evolution of the Schrödinger equation provides us with.[11]

Even this "precise simulation" is problematic to some considerable degree. We again have the issue of discrete approximation to a fundamentally

[10] Matters relevant to this issue are to be found in [44, 12, 46].

[11] The question may be raised that the seeming randomness that arises in chaotic classical dynamics might be the result of a deeper quantum-level actual randomness. However, this cannot be the full story, since quantum randomness also occurs with quantized classical systems that are not chaotic. Nevertheless, one may well speculate that in the non-linear modifications of quantum mechanics that I shall be later arguing for, such a connection between chaotic behaviour and the probabilistic aspects of present-day quantum theory could well be of relevance.

continuous mathematical model of reality. But with quantum systems there is also an additional problem confronting precise simulation, namely the vast size of the parameter space that is needed for the Schrödinger equation of a many-particle quantum system. This comes about because of the quantum entanglements referred to earlier. Every possible entanglement between individual particles of the system requires a separate complex-number parameter, so we require a parameter space that is exponentially large, in terms of the number of particles, and this rapidly becomes unmanageable if we are to keep track of everything that is going on. It may well be that the future development of *quantum computers* would find its main application in the simulation of quantum systems. We find in this collection, some discussion of the potential of quantum computers, though there no consensus is provided as to the likely future of this interesting area of developing technology.[12]

We see that despite the discreteness that has been introduced into physics via quantum mechanics, our present theories still require us to operate with real-number (or complex-number) functions rather than discrete ones. There are, however, proposals (e.g. [4]) in which the notion of "computation" is taken in a sense in which it applies *directly* to real-number operations, the real numbers that are employed in the physical theory being treated as real numbers, rather than, say, rational approximations to real numbers (such as finitely terminated binary or decimal approximations). In this way, simulations of physical processes can be carried out without resorting to approximations. This, however, can require that the initial data for a simulation be given as explicitly known functions, and that may not be realistic. Moreover, there are various *different* concepts of computability with real numbers, [41, 5, 33, 4, 28] which, unlike in the situation that arose for discrete (integer-valued) variables, where the Church-Turing concept appears to have provided a *single* generally accepted universal notion of "computation", there are many different proposals for real-number computability and no such generally accepted single version appears to be in evidence. Moreover, we unfortunately find that, according to a reasonable-looking notion of real-number computability, the action of the ordinary second-order wave operator turns out to be non-computable in certain circumstances (see e.g. Refs. [28, 29]). Whatever the ultimate verdict on real-number computability might be, it appears not to have settled down to something unambiguous as yet.

There is also the question of whether an exact theory of real-number

[12] Schmidhuber, Lloyd, Zurowski.

computability would have genuine relevance to how we model the physical world. Since our measurements of reality always contain some room for error—whether this be in a limit to the precision of a measurement or in a *probability* that a discrete parameter might take one or another value (as sometimes is the case with quantum mechanics)—it is unclear to me how such an exact theory of real-number computability might hold advantages over our present-day (Church–Turing) discrete-computational ideas. Although the present volume does not enter into a discussion of these matters, I do indeed believe that there are significant questions of importance here that should not be left aside (for example, see [29, 20, 41, 5]).

Several articles in this volume address the issue of whether, in some sense, the universe actually *is* a computer.[13] To me, this seems to be a somewhat strange idea. Although I can more-or-less understand what it might mean for it to be possible to have (theoretically) a computational *simulation* of all the actions of the physical universe,[14] which involves some sort of "constructivist" assumption[15] for the operation of the physical world, I find it much less clear what it might mean for the universe to *be* a computer. Various images come to mind, maybe suggested by how one chooses to picture a modern electronic computer in operation. Our picture might perhaps consist of a number of spatially separated "nodes" connected to one another by a system of "wires", where signals of some sort travel along the wires, and some clear-cut rules operate at the nodes, concerning what output is to arise for each possible input. There also needs to be some kind of direct access to an effectively unlimited storage area (this being an essential part of the Turing-machine aspirations of such a computer-like model). However, such a discrete picture and a fixed computer geometry does not very much resemble the standard present-day models that we have of the small-scale activity of the universe we inhabit. The discreteness of this picture is perhaps a little closer to some of the tentative proposals for a discrete physical universe, such as the "causal sets"[16] that I shall briefly return to later, which represent some attempts at radical ideas for what space-time might be "like" at the Planck scale.

Yet, there are some partial resemblances between such a computer-like picture and our (very well supported) present-day physical theories. These theories involve individual constituents, referred to as "quantum particles",

[13] Lloyd, Deutsch, Turner and Zuse.
[14] Bolognesi and Szudzik.
[15] Bauer.
[16] Bolognesi.

where each would have a classical-level description as being spatially "point-like"—though persisting in time, providing a classical space-time picture of a 1-dimensional "world-line". If these world-lines are to be thought of as the "wires" in the above computer-inspired picture, then the "nodes" could be thought of as the interaction places (or intersection points) between different particle world-lines. This would be not altogether unlike the computer image described above, though in standard theory, the topological geometry of the connections of nodes and wires would be part of the dynamics, and not fixed beforehand. Perhaps the lack of a fixed geometry of the connections would provide a picture more like the *amorphous* type of computer structure also considered in this volume,[17] than a conventional computer. However, it is still not clear how the "direct access to an effectively potentially unlimited storage area" is to be represented. More seriously, this is merely the *classical* picture that is conjured up by our descriptions of small-scale particle activity, where the quantum "picture" would consist (more or less) of a *superposition* of all these classical pictures, each weighted by a complex number. Such a "picture" perhaps gets a little closer to the way that a *quantum* computer[18] might be represented, but again there are the crucial issues raised by the topology of the connections being part of the dynamics and the absence of an "unlimited storage area", in the physical picture, which seem to me to represent fundamental differences between our universe picture and a quantum computer. In addition to all this, there is again the matter of how one treats the *continuum* in a computational way, which in quantum (field) theory is more properly the *complex* rather than the real continuum. Over-riding all this is the matter of how one actually gets information out of a quantum system. This requires an analysis of the *measurement problem* that I shall need to come to shortly.

I think that, all this notwithstanding, when people refer to the universe "being" a computer, the image that they have is not nearly so specific as anything like that suggested above. More likely, for our "computer universe" they might simply have in mind that not only can the universe's actions be precisely simulated in all its aspects, but that it has *no other* functional quality to it, distinct from this computational behaviour. More specifically, for our "computer universe" there would be likely to be some parameter $t$ (presumably a discrete one, which could be regarded as taking integer values)

[17] Hewitt, Teuscher, Margenstern and Wiedermann.

[18] Schmidhuber, Margenstern, Zurowski.

which is to describe the passage of time (not a very relativistic notion!), and the state of the universe at any one time (i.e. $t$-value) would have some computational description, and so could be completely encoded by a single natural number $S_t$. It would be the universe's job to compute $S_{t'}$ from $S_t$ whenever $t' > t$, and the universe would be considered to *be* a computer provided that not only is it able always to achieve this, but—more importantly—that this is the *sole* function of the universe. It seems to me if, on the other hand, the universe has any additional function, such as to assign a *reality* to any aspect of this description, then it would not simply be a computer, but it would be something more than this, succeeding in providing us with some kind of *ontology* that goes beyond the mere computational description.

To conclude this Foreword, I wish to present something that is much more in line with my *own* views as to the relation between computation and the nature of physical reality. To begin with, I should perhaps point out that my views have evolved considerably over the decades, but without much in the way of abrupt changes. Early on I had been of a fairly firm persuasion that there should be a discrete or combinatorial basis to physics, perhaps somewhat along the lines expressed in some of the articles[19] in this volume. In 1967 Erwin Kronheimer and I published a paper [14, 18] on the kind of causal sets referred to earlier in this Foreword, where the basic relationships between the elements are those of *causality*[20], mirroring the causal relations between events in continuous space-time, but where no continuity or smoothness is assumed, and where one could even envisage situations of this kind where the total number of these elements is finite.

Although I also had different reasons to be interested in spaces with a structure defined solely by causal relations—partly in view of their role in relation to singularity theorems [17, 11] (for the study of black holes and cosmology)—the causality relations not necessarily being tied to the notion of a smooth space-time manifold, I did not have much of an expectation that the true small-scale structure of our actual universe should be helpfully described in these terms. I had thought it much more probable that a different combinatorial idea, that I had been playing with a good deal earlier, namely that of *spin-networks* (see Ref. [19]) might have true relevance to the basis of physics (and indeed, much later, a version of spin-network theory was to

[19] Bolognesi, Schmidhuber, Lloyd, Wolfram, Zuse, Fredkin and Zenil.

[20] Bolognesi.

form part of the loop-variable approach to quantum gravity, [1] although the role that spin-networks acquire in loop-variable gravity is somewhat different from what I had originally envisaged).

Spin-network theory was based on one of the most striking parts of standard quantum mechanics, where a fundamental notion that is continuous in classical mechanics, is discrete in quantum mechanics, namely *angular momentum* (or *spin*). In fact, many of the most basic and counter-intuitive features of quantum mechanics, such as discreteness and (Bell) non-locality,[21] are most powerfully expressed in terms of quantum-mechanical spin. The puzzling relation between the continuous array of possibilities for the direction of a spin axis in our classical space-time pictures and the discrete (or "granular") nature of the quantum idea of "spin-axis direction" had always maintained a fundamental fascination for me. This and the basic non-locality of information in quantum mechanics come together in spin-network theory, where the classical idea of a spatial "direction" does not arise in a well-defined way until very large spin-network structures are present in order provide a good approximation to the continuous sphere of possible spatial directions. Specific mathematical devices for calculating the often extremely complicated expressions were developed, but everything remains completely discrete, and computational in the conventional sense of the word, continuity arising only in the limit of large numbers. The need to generalize the idea of spin-networks in order that the geometry of 4-dimensional space-time might be described, rather than just the sphere of spatial directions, finally found some satisfaction in the ideas of *twistor theory* (see Refs. [18, 23], Chapter 33). This provided a different way of looking at space-time geometry from what is usual—but now the idea of discreteness underlying the basis of physics began to fade, and became superseded by the magic of complex geometry and analysis.

One normally thinks of the space-time 4-manifold as being composed of "events" (i.e. space-time points), which are the basic elements of the geometry. Instead, twistor theory takes its basic elements to be modelled on entire histories of massless spinning particles in free flight. By a careful combination of ideas from space-time 4-geometry and the quantum-mechanical structure of relativistic angular momentum for massless particles, the concept of "twistor algebra" was developed. [18, 19] In special relativity, the basic concept of a *twistor*, which describes the kinematical structure of a spinning

[21] Breuer, Cabello, Schmidhuber and Zenil.

massless particle, finds its mathematical description as an element of the complex 4-vector space $\mathbb{T}$ referred to as "twistor space". The geometry of $\mathbb{T}$ relates to the real geometry of Minkowski 4-space $\mathbb{M}$ by means of an explicit geometrical correspondence, relating $\mathbb{M}$ directly to the complex 3-geometry of the *projective* twistor space $\mathbb{PT}$. It turns out that the complex numbers of quantum mechanics dovetail with those of the complex geometry of twistor theory in surprising ways, and that there is an intriguing interplay between the non-locality that naturally arises in the twistor description of quantum wavefunctions and the non-locality that we actually find in quantum phenomena. [24] In recent years, twistor theory has found considerable value in the calculation of high-energy scattering processes, where the rest-masses of the particles involved can be ignored, (See, for example, Ref. [2]) but many of the deeper issues confronting twistor theory remain unresolved.

It has always been an aim of twistor theory (still only partially fulfilled) that it should form a vehicle for the natural unification of quantum mechanics with general relativity. By this, I do not mean "quantum gravity" in the conventional sense in which this term is used. What is usually meant by quantum gravity is some scheme in which the ideas of Einstein's theory of gravity—namely *general relativity* (or else perhaps some modification of Einstein's theory)—is brought under the umbrella of quantum field theory. This viewpoint is to take the laws of quantum field theory as being *inviolate*, and that the ideas of general relativity must yield to those of quantum theory via some appropriate form of "quantization". My own view has always been different from this, as I believe that quantum theory itself, quite apart from its need to be unified with general relativity theory, is basically *self-inconsistent* and that some help is needed from *outside* the normal rules of quantum (field) theory. The view here is that the underlying principles of general relativity should help to supply this outside assistance.

This inconsistency is a very fundamental one, and is in a clear sense completely obvious (the "elephant in the room"!) as we shall see. As remarked upon earlier, we take the evolution of a quantum system in isolation to be governed by the Schrödinger equation—or, in more general terms, *unitary evolution*—and for which I use the symbol "U". But, as was remarked upon earlier, the reality of the world that we actually observe taking place about us tends *not* to be described directly by the solution $\Psi$ of this equation that we get by this U-evolution, but when an observation or "measurement" is deemed to have taken place, $\Psi$ is considered to "jump" to just one member $\Psi_r$ of a family of *superposed* alternative solutions

$$\Psi = \alpha_1\Psi_1 + \alpha_2\Psi_2 + \ldots + \alpha_n\Psi_n \tag{1}$$

where the respective squared moduli of the complex-number weightings $\alpha_1, \alpha_2, \ldots, \alpha_n$, supply the respective probabilities of each $\Psi_r$ being the result (the quantities $\Psi_r$ being assumed to be all normalized and mutually orthogonal). The "evolution process" whereby $\Psi$ is replaced by the particular $\Psi$r that happens to come about is the *reduction* of the state (collapse of the wavefunction) and I denote this process by the letter "R".[22]

Of course, there will be many such decompositions, for a given $\Psi$, depending on the choice of *basis* that is supposed to be determined by the choice of "measuring device". Indeed, we must allow that this measuring device is also part of the entire system under consideration, and so should have a quantum state that becomes entangled with the quantum system under examination. Nevertheless there is still taken to be a "jump" in the system as a whole as soon as the measurement is considered to have been made, where the different "pointer states" of the device are entangled with the different possible $\Psi_r$s that can result. It is *obvious* that this "jumping" from the state of the system (consisting of both the measuring device and system under examination, together with the entire relevant surrounding environment), from before measurement to after measurement, is normally not even continuous, let alone a solution of the Schrödinger equation: so R blatantly violates U (in almost all circumstances).

Why do physicists not normally consider this to be a contradiction in quantum mechanics? There are many responses, usually involving some subtle issue of "interpretation", according to which physicists try to circumvent this (seeming?) contradiction. Here is where the "many-worlds" viewpoint of Hugh Everett III is often invoked, [8, 6] whereby it is considered that *all* alternative outcomes simply(!) *co-exist* in quantum superposition, and that it is perhaps somehow a feature of our conscious perception processes that we always perceive only *one* of these alternatives. Despite this idea's popularity among many philosophically minded physicists (or physically well-educated philosophers), I find this viewpoint very unsatisfactory. I would agree that it is indeed where we are led, if we regard the U-process as inviolate, but to me this is to be taken as a *reductio ad absurdum* and a clear indication that we need to seek an improvement in current quantum mechanics. To put this

[22] In Von Neumann's classic book Mathematical Foundations of Quantum Mechanics [16], he introduced "R" and "U" under the respective names "process I" and "process II".

another way, even if the many-worlds viewpoint is in some sense "correct", it is still inadequate as a description of the physical world, for the simple reason that it does not, as it stands, describe the world that we actually observe, in which we find that something extremely well approximated by the R-process *actually* takes place when quantum superpositions of states that are sufficiently different from one another are involved.

What do I mean by "sufficiently different"? It is clear that mere *physical distance* apart, for the different material displacements involved in the superposed states, is not the correct criterion, because there have been well-confirmed experiments in which photon states tens of kilometres apart still maintain their quantum entanglements with one another, so that their various possible different polarization states remain in quantum superposition with each other even over such distances.[35] However, there are reasons to expect, from various foundational principles of Einstein's general theory of relativity, (see [7, 21, 25]) that when *mass displacements* between two quantum-superposed states get large, then such superpositions become unstable and ought to decay, in a roughly calculable time $\tau$, into one or the other, so that classical behaviour begins to take over from quantum behaviour. The estimate of $\tau$ is given by the formula

$$\tau \approx \frac{\hbar}{E_G} \tag{2}$$

where $E_G$ is the gravitational self-energy of the *difference* between the mass distributions in each of two quantum states under consideration, each being assumed to constitute a stationary state if on its own. Such a decay would represent a deviation from the standard *linearity* of U and might perhaps even be the result of some kind of chaotic behaviour arising in some non-linear generalization of present-day quantum mechanics. There are experiments currently under development that are aimed at testing this proposal, and we may perhaps anticipate results over the next several years (see Ref [15]).

For various reasons, partly concerned with the quantum *non-locality* referred to earlier (which only begins to present substantial problems for quantum realism when R is involved, treated as a *real* phenomenon), I would expect this change in current quantum mechanics to represent a *major revolution* and would not be at all easy to arrive at simply by "tinkering" with the Schrödinger equation. Indeed, my expectations are that such a theory would have to be *non-computable* in some very subtle way. Why am I mak-

ing such an assertion? The main reasons are rather convoluted, and I quite understand why some people regard my proposals as somewhat fanciful. Nevertheless, I am of the view that there is a good foundational rationale for a belief that something along these lines may actually be true!

The basic reason comes from Gödel's famous incompleteness theorems, which I regard as providing a strong case for human understanding being something essentially non-computable. The central argument is a familiar one, and I still find it difficult to comprehend why so many people are unwilling to take on board what would seem to be its fairly clear implication in this regard. In simple terms, the argument can be applied to our abilities to demonstrate the truth of certain mathematical propositions—which we can take to be of the form of $\Pi_1$-*sentences*. A $\Pi_1$-*sentence* is an assertion that some proposed Turing computation never terminates (examples being Wiles's "Fermat's last theorem" and Lagrange's theorem that every natural number is the sum of four squares). We might try to encapsulate, within some algorithmic procedure $\mathbb{A}$, all possible types of argument that can, in principle, be used to establish $\Pi_1$-*sentences*, according to human insight and understanding. This argument might be a *proof* within some given formal system $\mathbb{F}$, where $\mathbb{A}$ would be an algorithm for checking whether a proposed proof using the rules of $\mathbb{F}$ had been correctly carried out, giving the answer YES after a finite number of steps if this is indeed the case. What the Gödel(–Turing) theorem shows, in this context, is that if we have *trust* in $\mathbb{A}$ (and therefore in the soundness of such an $\mathbb{F}$, with regard to $\Pi_1$-*sentences*) that a "proved" $\Pi_1$-*sentence* is indeed *true* whenever $\mathbb{A}$ asserts YES, then one can explicitly exhibit a $\Pi_1$-*sentence* $G$ where our trust in $\mathbb{A}$ extends also to a trust in the truth of G, even though $\mathbb{A}$ itself is shown to be incapable of directly establishing $G$. In the case of Gödel's second incompleteness theorem, the $\Pi_1$-*sentence* $G(= G(\mathbb{F}))$ would be an assertion of the *consistency* of $\mathbb{F}$, and $G$ would be the $\Pi_1$-assertion that among the theorems of $\mathbb{F}$, there would be none whose negation is also a theorem of $\mathbb{F}$. Although our trust tells us that $\mathbb{A}$ would be unable to establish $G$ (i.e. the consistency of $\mathbb{F}$), our trust that $G$ is actually true follows from our *trust* in $\mathbb{A}$ (which *depends* on $\mathbb{F}$'s consistency—otherwise $\mathbb{F}$ would be able to establish $2 = 3$, a conclusion which we certainly would not trust). Our trust in the *use* of $\mathbb{F}$ as a means of establishing the truth of $\Pi_1$-*sentences* therefore carries us beyond the direct capabilities of $\mathbb{F}$, and enables us to assert that $G(\mathbb{F})$ is true, on the basis of that same trust, despite the fact that $\mathbb{F}$ does not contain G($\mathbb{F}$) among its theorems.

This is basically the thrust of Gödel's attack on formalism. Although the formalization of various areas of mathematics certainly has its value, allowing us to the transfer different aspects of human understanding and insight into computational procedures, Gödel shows us that these explicit procedures, once known—and *trusted*—cannot cover everything in mathematics that is accessible to understanding and insight.[23] And, indeed, this applies already for the relatively limited area of $\Pi_1$-*sentences*. Yet, a case can certainly be argued that this does *not* yet provide a demonstration that human insight is, at root, a non-algorithmic procedure, and I list here what appear to be the main arguments in support of that case, i.e. of criticisms of the above claim that the Gödel-type arguments show that human understanding is non-computational,

1. *Errors argument*—human mathematicians make errors, so rigorous Gödel-type arguments do not apply.

2. *Extreme complication argument*—the algorithms governing human mathematical understanding are so vastly complicated that their Gödel statements are completely beyond reach.

3. *Ignorance of the algorithm argument*—we do not know the algorithmic process underlying our mathematical understanding, so we cannot construct its Gödel statement.

I have tried to argue elsewhere [20] that (1), (2), and (3) do not invalidate the conclusion that our conscious understandings are very unlikely to be entirely the product of computational actions, and it is not my purpose to repeat such detailed arguments here. Nevertheless I briefly summarise my counter-arguments, in what follows.

The main point, with regard to (1) is that human errors are correctable. We are not so much concerned with the often erroneous gropings that mathematicians employ in their search for truth, but more the *ideals* that they grope *for* and, more importantly, measure their achievements against. It is their ability to *perceive* these ideals that we are concerned with, if only in principle, and it is this ability to perceive ideal mathematical truth that we are concerned with here, not the errors that we all make from time to time. (It may be evident from these comments that I *do* regard mathematical truth—especially with regard to matters so straight-forward as $\Pi_1$-*sentences*—as

[23] Zizzi.

something absolute, and external to ourselves. But I appreciate that others[24] are sometimes less sympathetic to this kind of viewpoint. I do not believe, however, that one's philosophical standpoint in this respect significantly affects the arguments that I am putting forward here.) With regard to (2) the point is somewhat similar. If the algorithms were *in principle* to be known, then their size or complication is of no real concern. This applies to a great many mathematical arguments. In Euclid's proof of the infinity of primes, for example, we need to consider primes that are so large that there would be no way to write them down explicitly in the entire observable universe, and to calculate the product of them all up to some such size in practice, is even more out of the question. But all this is irrelevant for the proof. Similar points apply to (2).

The argument (3) is, however, much more relevant to the discussion, and was basically Gödel's own reservation (referred to in the commentaries here[1]) with regard to making the strong conclusion that I am arguing for here. Rather than pushing the logical argument further, which is certainly possible to do (see Ref. [20]). I shall here merely indicate the extraordinary improbability of the needed algorithmic action arising in our heads, by the process of *natural selection*. Such an algorithm would have to have extraordinary sophistication, so as to be able to encapsulate, in its effective "formal system" many steps of "Gödelization". As an example, I have pointed out elsewhere [22] that whereas *Goodstein's theorem*, [9] whose meaning[2] is easily accessible even to those with little mathematical knowledge other than basic numerical notation, has been shown by Kirby and Paris [13] to be inaccessible by first-order Peano arithmetic (without a "Gödelization" step, that is), yet this theorem can be readily seen to be true through mathematical understanding. If our mathematical understanding is achieved by some (unknowable, but sound) algorithmic procedure, it would be a total mystery how it could have arisen through natural selection, when the experiences of our remote ancestors could have gained no benefit whatsoever from having such a sophisticated yet totally irrelevant algorithm planted in their brains!

If, then, it is accepted that our understanding of mathematics is not an algorithmic process, we must ask the question what kind of process can it be? A key issue, it seems to me, is that genuine *understanding* (at least in

[24] DeMol.
[1] Sieg.
[2] Velupillai.

our normal sense of this word) is something that requires awareness—as it would seem to me to be a misuse of the word "understanding" if it could be genuinely applied to an entity that had no actual *awareness* of the matter under its consideration. Awareness is the passive form of consciousness, so it seems to me that it was the evolutionary development of consciousness that is the key, and that such a quality could certainly have come about through natural selection, being able to confer an enormous selective advantage on those creatures possessing it. In saying this, I am expressing the view that consciousness is indeed *functional* and is not an "epiphenomenon" that simply happens to accompany certain kinds of cognitive processes. This view is certainly an implication of the quality of "understanding" requiring conscious awareness, since understanding is certainly functional.

I should make clear that I am making no claim to know—or to be able to define—what consciousness actually *is*, but its role in underlying "understanding" (whatever *that* is) seems to me to be of great evolutionary value, and could readily arise as a product of natural selection. I should also make clear that I am regarding the consciousness issue as a *scientific* one, and that I do *not* take the view that these are matters that are inaccessible to scientific investigation. I also take it that healthy wakeful human brains (as well as whatever other kind of animal brains may turn out also to be similarly capable) are able, somehow, to evoke consciousness by the application of those very same physical laws that are present throughout the universe, even though consciousness itself comes about only in the *very* special circumstances of organization that are needed to promote its appearance.

What kind of circumstance could that be, if we are asking for some sort of non-computable action to come about—when we bear in mind that the deterministic differential equations of classical or quantum physics seem to be of an essentially computable nature? My response to this query is that the non-computability must lie in hitherto undiscovered laws that could be of relevance here. (I am ignoring the issue, referred to earlier, of the discrete computational simulation of a continuous evolution. Yet, I do accept that there might be some questions of genuine relevance here that ought to be followed up more fully.) As far as I can see, the only big unknown, in physical laws, that could have genuine relevance here, is the U/R puzzle of quantum mechanics, referred to above. In almost all processes that take place, we have no need of the presumed New Theory that is to go beyond current quantum mechanics, mainly because its effects would go un-noticed, being swamped by the multifarious random influences of environmental decoherence. But, in

the brain, there might be relevant structures able to preserve quantum coherence up to a length of time at which the previously mentioned $\tau \approx \hbar/E_G$ criterion actually becomes relevant. Then, the normal purely probabilistic action that standard quantum theory's R-process provides us with is to be replaced by some subtle non-computational decision as to which choice the state reduction leads to. With a sophisticated brain organization, where the synaptic responses are sensitive to these choices, we can imagine that the output of the brain could indeed be usefully non-computational. This, indeed, is the basis of the "orchestrated objective reduction" (**Orch-OR**) scheme that Stuart Hameroff and I have proposed some years ago, where the above "relevant structures" would be *neuronal microtubules* of the appropriate type (see Refs. [10, 21, 27]).

It is hardly surprising that such a proposal has met with some considerable scepticism, mainly for the very understandable reason that to have body-temperature quantum coherence at anything like the level required is enormously far beyond the expectations of standard physical calculations applied to simplified models of cells. [34] Nevertheless, biological cells are, in fact, highly sophisticated structures,[3] and one may reasonably expect that, when the structures of certain cell parts are dedicated in the appropriate directions, their behaviour might exhibit quite unusual quantum-mechanical properties.[4] In fact, recent experiments carried out in Japan by Anirban Bandyopadhyay [3] and his co-workers appear to have demonstrated that highly intriguing quantum-coherent effects *do* actually take place in body-temperature neuronal microtubules. These results are, as of now, preliminary, but they do appear to provide some encouragement for the **Orch-OR** scheme, and it will be very interesting to see how things develop.

Even if all of this is accepted, we may still ask what would be the use of a little bit of non-computable action, from time to time, for the operation of the brain? Indeed, there would not be much value in this unless the quantum coherence is of a very global character, involving large areas of the brain, and the process would have to act in some globally coherent way. This is indeed the **Orch-OR** picture, and we take it that moments of consciousness occur when state reduction occurs at many sites (in microtubules) at once in an *orchestrated* way, so that the synapse strengths are influenced in many places and a concerted influence results, as would be expected for conscious actions.

[3] Margenstern, Ehrenfeucht et al., Rozenberg and Zenil.

[4] Zizzi.

The results of particular acts of conscious understanding would be unlikely to be usually anything simple, and would depend upon the experience of memories as well as on logic. But the non-computable ingredient is taken to be essential, for the reasons described above. According to this view, our conscious actions are calling upon parts of physics—encompassed in a New Theory that is presently unknown in detail. The impact of this theory on processes *not* organized in this way would not be evident. But it would make its mark on systems—such as wakeful healthy human brains—where it emerges as conscious actions and perceptions. The non-computable effects of this New Theory would emerge in this way and result in actions that are described as "hypercomputational".

How far outside the normal scheme of computational physics would these hypercomputational actions be? Since the Gödelian insight that allows us to transcend a given trusted formal system $\mathbb{F}$ provides this insight in the form of a $\Pi_1$-*sentence*, namely $G(\mathbb{F})$, we might expect that we could model such hypercomputational actions in the form of a Turing *oracle-machine*,[5] where the oracle is able to assert the truth or falsity of $\Pi_1$-*sentence*. However this would not be sufficient (nor does it appear to be necessary), as we can apply a Gödel-type "diagonalization" insight again on $\Pi_1$-*sentence*-oracle machines to transcend these devices also. In a recent article, [26] I consider a type of oracle that I refer to as a "cautious oracle", which is intended to model a little more closely the kind of thing that one might consider idealized human mathematicians might be capable of, where the cautious oracle can examine a $\Pi_n$-*sentence* (for any natural number $n$) and either respond "true" or "false" (necessarily truthfully in each case), or else confess to being unable to supply an answer or, failing any of these, simply continue pondering indefinitely without ever providing an answer at all. Again a Gödel-type diagonalization allows us the insight to transcend any such a device's capabilities! Whatever kind of hypercomputational capabilities such a "New Theory" might confer, it appears to be something very subtle. It is some sort of never-ending capability of being able to "stand back" and contemplate whatever structure had been considered previously. This seems to be a quality that consciousness is able to achieve, but how one incorporates this kind of thing into a physical theory is hard to imagine, as our present-day theories stand.

[5] Chaitin, Dershowitz.

# References


[1] Ashtekar, A. and Lewandowski, J. Background independent quantum gravity: a status report. *Class. Quantum Grav. 21,* R53.[gr-qc/0404018], 2004.

[2] Arkani-Hamed, N., Cachazo, F., Cheung, C., and Kaplan, J. The S-Matrix in Twistor Space. arXiv:0903.2110v2 [hep-th], 2009

[3] Sahu, S., Ghosh, S., Hirata, K., Fujita, D. and Bandyopadhyay, A. Ultrafast microtubule growth through radio-frequency-induced resonant excitation of tubulin and small-molecule drugs, to appear in Nature Materials.

[4] Blum, L. Alan Turing and the Other Theory of Computation (on Turing's Rounding-off Errors in Matrix Processes) in S Barry Cooper and Jan van Leeuwen (eds), Alan Turing – *His Work and Impact.* Elsevier, 2012.

[5] Bridges, D.S. Can constructive mathematics be applied in physics? *J. Philos. Logic 28,* 439–53, 1999.

[6] DeWitt, B.S. and Graham, R.D., eds. *The Many-Worlds Interpretation of Quantum Mechanics.* Princeton Univ. Press, Princeton, 1973.

[7] Diósi, L. Models for universal reduction of macroscopic quantum fluctuations. *Phys. Rev. A40,* 1165–74, 1989.

[8] Everett, H. "Relative State" formulation of quantum mechanics. In J.A. Wheeler and W.H. Zurek (eds), *Quantum Theory and Measurement.* Princeton Univ. Press, Princeton, 1983), originally in *Revs. of Modern Physics, 29*, 454–62, 1957.

[9] Goodstein, R.L. On the restricted ordinal theorem. *J. Symbolic Logic 9,* 33–41, 1944.

[10] Hameroff, S.R. and Penrose, R. Conscious events as orchestrated space-time selections. *J. Consciousness Studies 3,* 36–63, 1996.

[11] Hawking, S.W. and Penrose, R. The singularities of gravitational collapse and cosmology, *Proc. Roy. Soc.,* London, A314, 529–548, 1970.

[12] Israeli, N. and Goldenfeld, N. Computational irreducibility and the predictability of complex physical systems. *Phys. Rev. Lett. 92,* 074105, 2004.

[13] Kirby, L.A.S. and Paris, J.B. Accessible independence results for Peano arithmetic, *Bull. L.ond. Math. Soc. 14,* 285–93, 1982.

[14] Kronheimer E.H. and Penrose, R. On the structure of causal spaces. *Proc. Camb. Phil. Soc. 63,* 481–501, 1967.

[15] Marshall, W., Simon, C., Penrose, R. and Bouwmeester, D. Towards Quantum Superpositions of a Mirror, *Phys. Rev. Lett.,* Vol. 91, Issue 13, 2003.

[16] von Neumann, J. *Mathematical Foundations of Quantum Mechanics.* (Princeton Univ. Press, Princeton), 1955.

[17] Penrose, R. Gravitational collapse and space-time singularities, *Phys. Rev. Lett. 14,* 57–59, 1965.

[18] Penrose, R. Twistor algebra, *J. Math. Phys. 8,* 345–66, 1967.

[19] Penrose, R. Angular momentum: an approach to combinatorial space-time. In T. Bastin (ed), *Quantum theory and Beyond,* Cambridge University Press, Cambridge, 1971.

[20] Penrose, R. *Shadows of the Mind; An Approach to the Missing Science of Consciousness*, Oxford Univ. Press, Oxford, 1994.

[21] Penrose, R. On gravity's role in quantum state reduction. *Gen. Rel. Grav. 28,* 581–600, 1996.

[22] Penrose, R. Can a computer understand? In Rose, S. (ed), *From Brains to Consciousness? Essays on the New Sciences of the Mind,* Allen Lane, The Penguin Press, London) 154–179, 1998.

[23] Penrose, R. *The Road to Reality: A Complete Guide to the Laws of the Universe*, Jonathan Cape, London, 2004.

[24] Penrose, R. The twistor approach to space-time structures. In A. Ashtekar (ed), *100 Years of Relativity; Space-time Structure: Einstein and Beyond*, World Scientific, Singapore, 2005.

[25] Penrose, R. Black holes, quantum theory and cosmology (Fourth International Workshop DICE 2008) J. Physics, *Conf. Ser. 174*, 012001. doi: 10.1088/1742-6596/174/1/012001, 2009.

[26] Penrose, R. On attempting to model the mathematical mind, in Cooper, B.S. and Hodges, A. (eds), *The Alan Turing Year - The Once and Future Turing*, Cambridge University Press, 2012.

[27] Penrose, R. and Hameroff, S. Consciousness in the Universe: Neuroscience, Quantum Space-Time Geometry and **Orch OR** Theory. *Journal of Cosmology, Vol. 14,* 2011.

[28] Pour-El, M.B. and Richards, I. The wave equation with computable initial data such that its unique solution is not computable, Adv. in Math. 39, 215–239, 1981.

[29] Pour-El, M.B. and Richards, I. Computability in Analysis and Physics. *Perspect. Math. Logic*, (Springer-Verlag, Berlin, Heidelberg), 206 pp., 2003.

[30] Russell, B. *The Analysis of Matter* (Allen and Unwin; reprinted 1954, Dover Publ. Inc., New York), 1927

[31] Sakharov, A.D. Vacuum Quantum Fluctuations in Curved Space and The Theory of Gravitation, *Sov. Phys. Dokl., 12,* 1040 [Dokl. Akad. Nauk Ser. Fiz. 177, 70]. Reprinted: (2000) *Gen. Rel. Grav., 32* 365–367, 1968.

[32] Schrödinger, E. *Science and Humanism: Physics in Our Time.* (Cambridge Univ. Press, Cambridge), 1952.

[33] Stannett, M. Computation and Hypercomputation. *Minds and Machines,13*, 115–53, 2003.

[34] Tegmark, M. "Importance of quantum coherence in brain processes," *Phys. Rev. E,* 61, pp. 4194–4206, 2000.

[35] Tittel, W., Brendel, J., Gisin, B., Herzog, T., Zbinden, H., and Gisin N. Experimental demonstration of quantum-correlations over more than 10 kilometers arXiv:quant-ph/9707042v3, 2008.

[36] Turing, A.M. On computable numbers, with an application to the Entscheidungsproblem, *Proc. Lond. Math. Soc. (ser. 2) 42,* 230–265; a correction 43, 544–546, 1937.

[37] Turing, A.M. Computability and 1-definability. *J. Symb. Log., 2,* 153–163, 1937.

[38] Turing, A.M. Systems of logic based on ordinals. *P. Lond. Math. Soc., 45 (2),* 161–228, 1939.

[39] Turing, A.M. 'Intelligent machinery', with AMS corrections and additions. Pages numbered 1–37, with 2 un-numbered pages of references and notes. Page 1 has MS note by R.O. Gandy, 'Turing's typed draft'.n.d., 1948.

[40] Turing, A.M. Computing machinery and intelligence, *Mind 59* no. 236; reprinted in D.R. Hofstadter and D.C. Dennett (eds), *The Mind's I*, Basic Books, Inc.; Penguin Books, Ltd; Harmondsworth, Middx. 1981, 1950.

[41] Weihrauch, K. *Computable Analysis: An Introduction.* Texts in Theoretical Computer Science, Springer, 2000.

[42] Wheeler, J.A. *Geometrodynamics* (Società Italiana Fisica: Questioni di fisica moderna, V.1, and Academic Press, Inc., New York), 1982.

[43] Wheeler, J.A. Law without law, in Quantum Theory and Measurement. In J.A. Wheeler, J.A. and Zurek, W.H. (eds), Princeton Univ. Press, Princeton, pp. 182–213, 1983.

[44] Wolfram, S. *A New Kind of Science,* Wolfram Media Inc, 2002.

[45] Zenil, H. and Delahaye, J.-P. On the Algorithmic Nature of the World. In G. Dodig-Crnkovic and M. Burgin (eds), *Information and Computation,* World Scientific, 2010.

[46] Zenil, H, Soler-Toscano, F., and Joosten, J. J. Empirical Encounters With Computational Irreducibility and Unpredictability, *Minds and Machines, vol. 21,* 2011.

arXiv:1211.1374v1 [gr-qc] 6 Nov 2012

# Testing General Relativity and gravitational physics using the LARES satellite

Ignazio Ciufolini[1], Antonio Paolozzi[2], Erricos Pavlis[3], John Ries[4], Vahe Gurzadyan[5], Rolf Koenig[6], Richard Matzner[7], Roger Penrose[8], Giampiero Sindoni[2]

*1.Dipartimento di Ingegneria dell'Innovazione, University of Salento, and INFN, Lecce, Italy,*
*2.Scuola di Ingegneria Aerospaziale and DIAEE, Sapienza Università, di Roma. Rome. Italy,*
*3.Planetary Geodynamics Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD, USA,*
*4.Center for Space Research, The University of Texas at Austin, TX, USA,*
*5.Center for Cosmology and Astrophysics, Alikhanian National Laboratory, Yerevan, Armenia,*
*6.Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences,*
*7.Center for Relativity, The University of Texas at Austin, TX, USA,*
*8.Mathematical Institute, University of Oxford, UK*

The discovery of the accelerating expansion of the Universe, thought to be driven by a mysterious form of 'dark energy' constituting most of the Universe, has further revived the interest in testing Einstein's theory of General Relativity. At the very foundation of Einstein's theory is the geodesic motion of a small, structureless test-particle. Depending on the physical context, a star, planet or satellite can behave very nearly like a test-particle, so geodesic motion is used to calculate the advance of the perihelion of a planet's orbit, the dynamics of a binary pulsar system and of an Earth orbiting satellite. Verifying geodesic motion is then a test of paramount importance to General Relativity and other theories of fundamental physics. On the basis of the first few months of observations of the recently launched satellite LARES, its orbit shows the best agreement of any satellite with the test-particle motion predicted by General Relativity. That is, after modelling its known non-gravitational perturbations, the LARES orbit shows the smallest deviations from geodesic motion of any artificial satellite: its residual mean acceleration away from geodesic motion is less than $4 \times 10^{-12} m/s^2$. LARES-type satellites can thus be used for accurate measurements and for tests of gravitational and fundamental physics. Already with only a few months of observation, LARES provides smaller scatter in the determination of several low-degree geopotential coefficients (Earth gravitational deviations from sphericity) than available from observations of any other satellite or combination of satellites.

## 1. Introduction

General Relativity is a fundamental concept for understanding the universe that we observe.[1-3] It describes one of the four fundamental interactions of nature, the gravitational interaction governing the dynamics of large-scale systems and bodies such as planets, stars, galaxies and the Universe, as well as our daily attraction towards the center of Earth. Einstein's theory is a very well verified description of gravity.[3-5] But it has encountered unexpected developments in observational

cosmology. The study of distant supernovae led to a discovery that distant galaxies accelerate away from us.[6,7] Since then, this dark energy,[8,9] regarded as a new exotic physical substance that is accelerating the expansion of the Universe, is at the center of attention of physics. And careful studies of that expansion also imply the presence of a large fraction of invisible (dark) but normally attracting matter,[10] dark energy and dark matter together constitute approximately 96% of the mass-energy of the Universe.

Rather than assuming the existence of these exotic material sources, the accelerated expansion of the Universe might be explained by changing the fundamental description of gravity. Some theories of gravitation alternative to General Relativity can explain the acceleration,[11,12] others have been ruled out.[13] Because most such theories have been much less intensively studied than General Relativity, continuing careful experimental tests of all aspects of gravitational interaction are needed.

General Relativity explains the gravitational interaction as the curvature of spacetime generated by mass-energy and mass-energy currents via the Einstein field equations.[1–3] For example, the gravitational attraction of Earth on its Moon and artificial satellites is explained by General Relativity via the spacetime curvature generated by the Earth's mass. The motion of any test-body within the gravitational field of another massive body, e.g., the motion of a 'small' satellite around Earth, is simply determined by a geodesic of spacetime with curvature generated by the massive body. The Moon and the Earth artificial satellites approximately follow geodesics of spacetime with orbital perturbations from an ideal geodesic path due their finite size and the non-gravitational forces acting on them. A timelike geodesic path (world-line) in spacetime's Lorentzian geometry is one that locally maximizes proper time, in analogy with the length-minimizing property of Euclidean straight lines. A test-particle that follows geodesic motion is an electrically neutral body, with negligible gravitational binding energy compared to its rest mass, negligible angular momentum and is small enough that the inhomogeneities of the gravitational field within its volume have a negligible effect on its motion. Furthermore, non-gravitational perturbations must not influence its motion.

Thus, geodesic motion is at the foundation of General Relativity and of any other theory where the gravitational interaction is described by spacetime curvature dynamically generated by mass-energy. Therefore, the creation of the best possible approximation for the free motion of a test particle, a spacetime geodesic, is a profound goal for experiments dedicated to the study of the spacetime geometry in the vicinity of a body, yielding high precision tests of General Relativity and constraints on alternative gravitational theories.

Important issues are now addressed regarding the approximation to a geodesic that is provided by the motion of an actually extended body. In General Relativity,[14,15] the problem of an extended body is subtle, due not only to the non-linearity of the equations of motion, but also to the need to deal with the internal structure of the compact body, constructed of continuous media, where kinetic variables and thermodynamic potentials are involved. Further, there may be intrinsically non-

local effects arising from the internal structure of the extended body, such as tidal influences. Moreover, there are problems concerning the approximations that need to be made in order to describe a given extended body as a test particle moving along a geodesic. These problems are related to the fact that many of the common Newtonian gravitational concepts such as the 'center of mass', 'total mass' or 'size' of an extended material body do not have well defined counterparts in General Relativity.[16]

The Ehlers-Geroch theorem[17] (generalizing the result in[18]) attributes a geodesic $\gamma$ to the trajectory of an extended body with a small enough own gravitational field, if for a Lorentzian metric the Einstein tensor satisfies the so-called dominant energy condition,[2] this tensor being non-zero in some neighborhood of $\gamma$ and vanishing at its boundaries.[19] This theorem, asserting that *'small massive bodies move on near-geodesics'*, thus achieves a rigorous bridge from General Relativity to space experiments with 'small' satellites which suggests a high level of suppression of non-gravitational and self-gravitational effects from the satellite's own small gravitational field. This enables us to consider the satellite's motion to be nearly geodesic and hence provides a genuine testing ground for General Relativity's effects.

Given the extreme weakness of the gravitational interaction with respect to the other interactions of nature, the space environment is the ideal laboratory to test gravitational and fundamental physics. However, in order to test gravitational physics, a satellite must behave as nearly as possible as a test-particle and must be as little as possible affected by non-gravitational perturbations such as radiation pressure and atmospheric drag. In addition, its position must be determined with extreme accuracy.

The best realization of an orbiting test particle is LARES (LAser RElativity Satellite), Figure 1, which was successfully launched on February $13^{th}$ 2012 into a medium altitude, circular Earth orbit.[20] Its position can be measured with very high accuracy using the Satellite Laser Ranging (SLR) technique. Short-duration laser pulses (with a typical width of 10 ps for the state-of-the-art systems) are emitted from lasers on Earth and then reflected back by retro-reflectors on the artificial satellites. By measuring the total round-trip travel time, it is possible to determine the instantaneous distance to the satellite with an accuracy of a few millimeters. There are a number of spherical laser ranged artificial satellites, among which are Starlette, Stella, LAGEOS (LAser GEOdynamics Satellite) and LAGEOS-2, Ajisai, Etalon-1 and Etalon-2. The tracking data collected by the SLR network, illustrated in Figure 2, are analyzed, organized and distributed by the International Laser Ranging Service (ILRS).[21]

However, in order to test gravitational physics, we not only need to measure the position of a body with extreme accuracy, but we also need it to behave like a test-particle. In space, a test-particle can be realized in two ways: a small drag-free satellite or a small spacecraft with high density and an extremely small area to mass ratio.

A drag-free satellite is an active spacecraft designed to reduce its non-

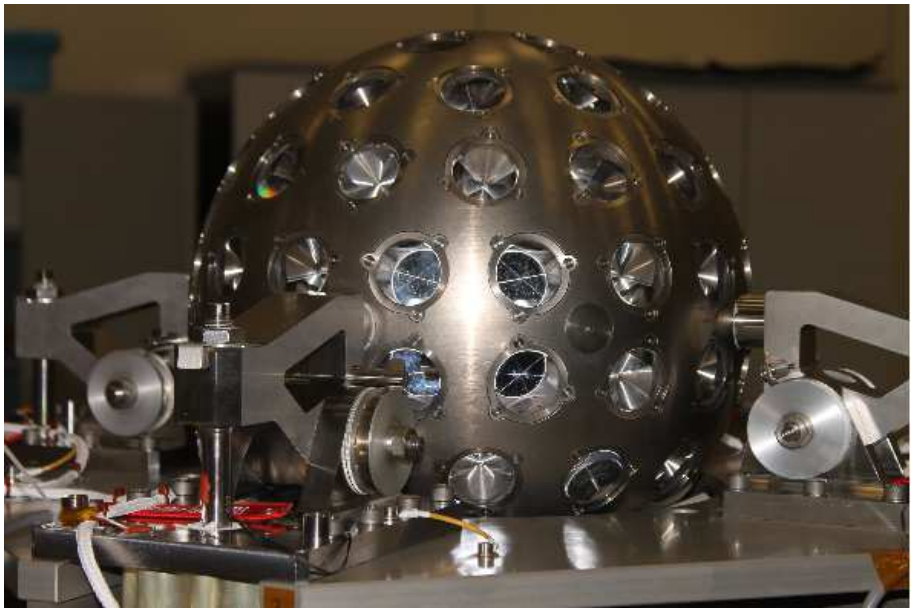


Fig. 1. The LARES satellite on the separation system (courtesy of ASI).

gravitational orbital perturbations to extremely small values. It can be realized using a proof mass placed close to the satellite's center of mass, in a cavity that completely shields the proof mass. The satellite is then controlled by thrusters to follow the proof mass in such a way that the distance between the proof mass and the inner walls of the cavity does not change.Since the cavity is closed, the spherical proof mass is shielded from non-gravitational perturbations, such as radiation pressure and atmospheric drag. In the ideal case when the effects of other disturbing forces are negligible, the orbit of the proof mass, and therefore of the satellite itself, is determined only by gravitational forces. The only disturbing forces that can act on the proof mass arise from the satellite itself or from an interaction that can penetrate the shield. In the case of the Gravity Probe-B satellite, a mean residual acceleration of about $40 \times 10^{-12} m/s^2$ was achieved.[22]

For a passive satellite (with no drag-free system), the key characteristic that determines the level of attenuation of the non-gravitational perturbations is the density, reflected by the ratio between its cross-sectional area and its mass. Until the launch of LARES, the two LAGEOS (LAser GEOdynamics Satellite) satellites had the smallest ratio of cross-sectional area to mass of any other artificial satellite[15] and were the best available test-particles. LAGEOS and LAGEOS-2 have an essentially identical structure but different orbits. They are passive, spherical satellites covered with retro-reflectors and made of heavy brass and aluminum. The mass of LAGEOS is 407 kg and that of LAGEOS-2 is 405.4 kg; their radius is 30 cm. The ratio of cross-sectional area to mass is approximately $0.0007 m^2/kg$ for both satellites. LAGEOS

was launched in 1976 by NASA, and LAGEOS-2 by NASA and ASI (Italian Space Agency) in 1992. The semi-major axis of LAGEOS' orbit is approximately 12,270 km, the period is 3.76 h, the eccentricity is 0.004 and the inclination is $109.8^o$. LAGEOS-2 is at a similar altitude but the orbit eccentricity is larger (0.014) and the inclination is lower ($52.6^o$).

LARES (LAser RElativity Satellite), also from ASI, was launched with the qualification flight of VEGA, the new European Space Agency's (ESA's) launch vehicle developed by ELV (Avio-ASI). LARES is a spherical laser ranged satellite covered with 92 retro-reflectors with a radius of 18.2 cm. It is made of a tungsten alloy, with a total mass of 386.8 kg,[20] resulting in a cross-sectional area to mass ratio that is about 2.6 times smaller than LAGEOS. It has almost certainly the highest mean density of any object orbiting in the Solar System. It is currently very well observed by the ILRS stations that are located all over the world (Figure 2). The LARES orbital elements are: semi-major axis 7820 km, orbital eccentricity 0.0007 and orbital inclination $69.5^o$.

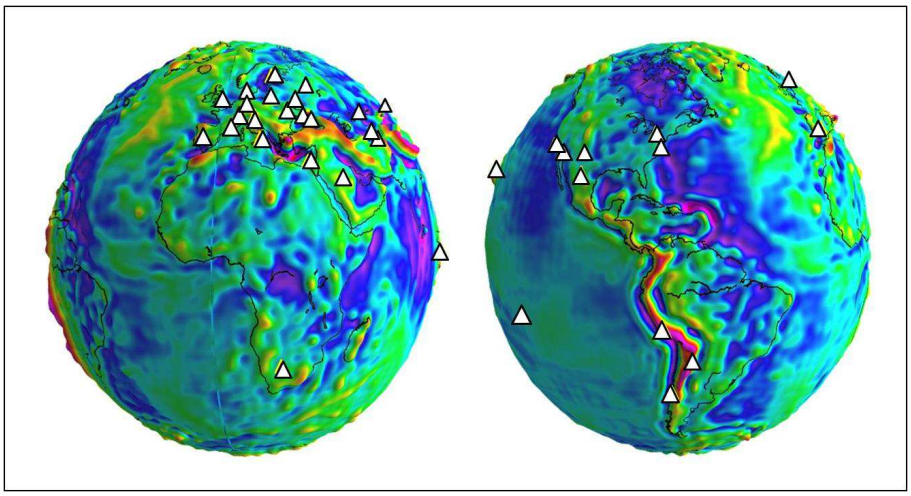


Fig. 2. The current International Laser Ranging Service networkshown on a model of Earth's gravitational field obtained by the GRACE space mission[24,25]

The extremely small cross-sectional area to mass ratio of LARES, which is smaller than that of any other artificial satellite, and its special structure, a solid-sphere with high thermal conductivity, ensure that the non-gravitational orbit perturbations are smaller than for any other satellite, in spite of its lower altitude compared to LAGEOS. This behavior has been experimentally confirmed using the first few months of laser ranging observations.

We processed the LARES laser ranging data based on the first seven 15-day arcs using the orbital analysis and data reductionsystems UTOPIA of UT/CSR (Center for Space Research of The University of Texas at Austin), GEODYN II

of NASA Goddard, and EPOS-OC of GFZ (Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences). In one case using GEODYN, we estimated only a single residual along-track empirical acceleration, and in another case using UTOPIA, together with the residual acceleration, we estimated the degree-2 and -3 geopotential coefficients. The latter case was intended to take into account the lower altitude of LARES and thus the larger effect of gravity model errors on the orbit; the estimation of the gravity terms had little effect on the residual acceleration. For LARES, we modeled drag using its 'best fit' drag coefficient $C_d$. The reflectivity coefficient, $C_\gamma$, was also the 'best fit' value, which was found to be consistent with the measured surface properties and with similar spherical satellites. In all cases, state-of-the art satellite orbit dynamical models were employed, including all the general relativistic post-Newtonian corrections, GRACE-based mean gravity field models,[24,25] modern models for the ocean and solid earth tides, as well as solar radiation pressure, Earth albedo and atmospheric drag.[26,28] No 'thermal thrust'[29] models (described more fully below) were used. For the 105 days analyzed, both GEODYN and UTOPIA determined that the residual Pavlisalong-track accelerations for LARES were only about $0.4 \times 10^{-12} m/s^2$, whereas for the two LAGEOS satellites, the acceleration residuals were $1-2\times10^{-12} m/s^2$. As a further test, we fit the drag coefficient $C_d$ of LARES over the odd 15-days arcs and we then applied the mean $C_d$ obtained over these odd arcs to estimate the mean residual acceleration over the even arcs and vice-versa. We found again a residual acceleration of LARES of less than $0.4 \times 10^{-12} m/s^2$. This is particularly impressive given that LARES is far lower in the Earth's atmosphere than LAGEOS.

The residual along-track accelerations of a satellite provide a measure of the level of suppression of its non-gravitational perturbations: atmospheric drag, solar and terrestrial radiation pressure and thermal-thrust effects. Atmospheric drag acts primarily along the satellite's velocity vector, while solar radiation pressure, terrestrial radiation pressure (the visible and infrared radiation from Earth) and thermal-thrust effects will all have some contribution along-track as well.The 'classic' Yarkovsky effect on a spinning satellite is a thermal thrust resulting from the anisotropic temperature distribution over the satellite's surface caused by solar heating. In particular, a variation of this effect due to the Earth's infrared radiation is the Earth-Yarkovsky or Yarkovsky-Rubincam effect.[29] Infrared radiation from Earth is absorbed by the retro-reflectors; due to their thermal inertia and the rotation of the satellite, a latitudinal temperature gradient develops. The corresponding thermal radiation causes a significant along-track acceleration in the direction opposite to the satellite's motion. For LAGEOS and LAGEOS-2, this is estimated to be of the order of $1 \times 10^{-12} m/s^2$. In addition to the Earth-Yarkovsky effect, there are also solar-Yarkovsky and asymmetric reflectivity effects that further perturb the LAGEOS orbits;[30] these also tend to be on the order of $1 - 2 \times 10^{-12} m/s^2$.

One of the reasons for the smaller residual acceleration of LARES with respect to LAGEOS, in spite of the lower altitude of LARES, is that the various thermal thrust effects should be much smaller on LARES than on LAGEOS. Indeed,

LARES is much smaller than LAGEOS (18 cm radius versus 30 cm radius for LAGEOS), and it has higher thermal conductivity since it is a solid one-piece sphere. In contrast, LAGEOS is constructed from three separate pieces, thus decreasing the overall thermal conductivity properties. Furthermore, the effect of the thermal acceleration due to the retro-reflectors (which are the main source of the Earth and solar Yarkovsky effects) is also smaller because of the smaller cross-sectional-area to mass ratio of LARES and because the total surface area of the retro-reflectors is smaller on LARES (about 26% of the total surface area) than on LAGEOS (about 43%).

The effects of the residual unmodelled along-track acceleration on the orbits of LARES and LAGEOS are illustrated in Figure 3A where we plot the change in the distance from their 'ideal' orbit caused by the unmodelled along-track accelerations. In Figure 3B, we show the effects of the residual unmodelled along-track acceleration on the orbits of LARES, LAGEOS and Starlette. The axis of ordinates may be thought of as representing an 'ideal' reference world line of LARES, LAGEOS and Starlette, 'ideal' in the sense that all of its orbital perturbations are known. Figures 3A and 3B show the unmodelled deviations from geodesic motion for LARES, LAGEOS and Starlette (once the known non-gravitational perturbations are removed, to the extent permitted by our current models) due to the unmodelled along-track accelerations. In these figures, we show the effect of a typical residual unmodelled along-track acceleration of $1\times10^{-12}m/s^2$ for LAGEOS, $0.4\times10^{-12}m/s^2$ for LARES and $40\times10^{-12}m/s^2$ for Starlette. Since all the general relativistic post-Newtonian corrections were included in our orbital analyses, these figures show the level of agreement of the LARES and LAGEOS orbits with the geodesic motion predicted by General Relativity.

It must be stressed that a residual unmodelled out-of-plane acceleration of the order of magnitude of the unmodelled along-track acceleration observed on LARES will produce an extremely small *secular* variation of the longitude of its node (the intersection of a satellite's orbital plane with the Earth's equatorial plane), i.e., of its orbital angular momentum. For example, by considering an out-of-plane acceleration with amplitude of $0.4\times10^{-12}m/s^2$, its effect on the node of LARES would be many orders of magnitude smaller than the tiny secular drift of the node of LARES due to frame-dragging[31] of about 118 milliarcsec/yr. Therefore, LARES, together with the LAGEOS satellites, and with the determination of Earth's gravity field obtained by the GRACE mission, will be used to accurately measure the frame-dragging effect predicted by General Relativity, improving by perhaps an order of magnitude the accuracy of previous frame-dragging measurements by the LAGEOS satellites.[32,33]

In fundamental physics, limits on certain possible low-energy consequences of string theory that may be related to dark energy and quintessence have been set using measurements of frame-dragging by the LAGEOS satellites.[34] These limits will be improved with LARES. In addition to these fundamental physics tests, the LARES satellite will be an important geodetic contribution to the monitoring of the long-wavelength variations in Earth's gravity field due to mass redistribution due to

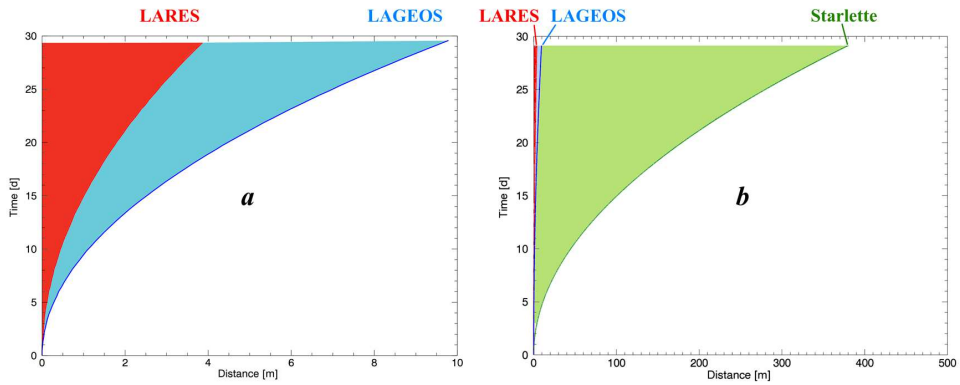



Fig. 3. A. The red curve represents the change of distance between a "test-particle" following a spacetime geodesic, represented here by the axis of ordinates in a frame co-moving with the test-particle, and a similar particle perturbed by the average unmodelled along-track acceleration of the magnitude observed on the LARES satellite of approximately $0.4 \times 10^{-12} m/s^2$. The blue curve represents the change of distance between a test-particle and a similar particle perturbed by an average along-track acceleration of the typical size of the unmodelled along-track acceleration observed on the LAGEOS satellites of the order of $1 \times 10^{-12} m/s^2$. The axis of ordinates may be thought of to represent a spacetime geodesic followed by LARES or LAGEOS after removing all the known and unmodelled non-gravitational perturbations. Fig.3 B. Similar plots for LARES, LAGEOS and Starlette are shown. For Starlette, we used a typical residual acceleration of the order of $40 \times 10^{-12} m/s^2$

the effects of global climate change. Until there are continuous GRACE-type gravity-monitoring missions, the SLR tracking to the 'cannonball' satellites continue to provide a critical measure of the long-term long-wavelength gravity changes. Figure 4 illustrates the improved observability that LARES provides for the low-degree (long-wavelength) gravity terms (the degree 2 and 3 terms mentioned previously). The increased error with increasing degree reflects the signal attenuation that is a function of the satellite altitude, which is more severe for the higher LAGEOS satellites than for the lower LARES and Starlette satellites. While LARES is only slightly higher than Starlette (the Starlette altitude ranges from 800 to 1100 km compared to about 1440 km for LARES), LARES is able to resolve the low degree terms as well or better than Starlette, and much better than the LAGEOS satellites. In particular, for $C_{30}$, where the biasing effect of the thermal-thrust forces becomes important, LARES is considerably better than LAGEOS or LAGEOS-2.

## Acknowledgments

In conclusion, LARES provides the best available test-particle in the Solar System for tests of gravitational physics and General Relativity, e.g., for the accurate mea-

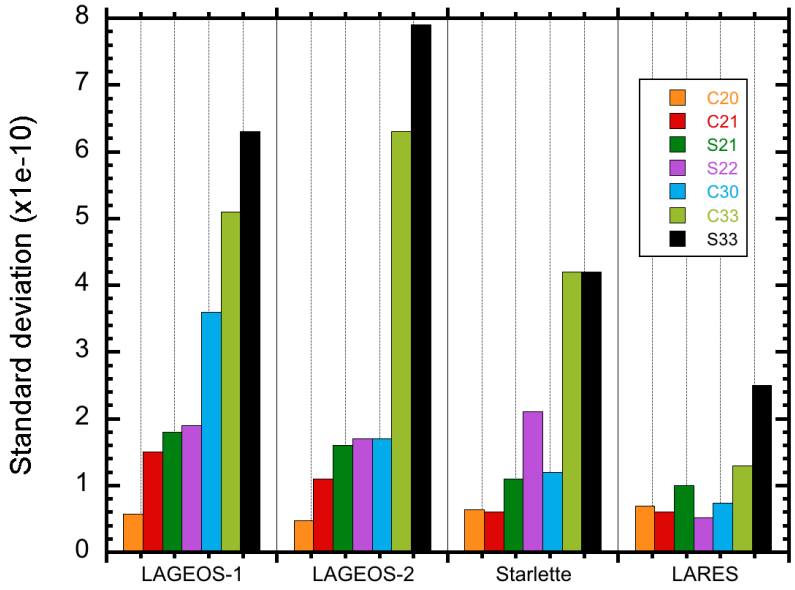



Fig. 4. Statistics of the scatter in the determination of several low-degree geopotential coefficientsusing LAGEOS, LAGEOS-2, Starlette and LARES.

surement of frame-dragging and,after modelling its known non-gravitational perturbations,its orbit shows the best agreement of any satellitewith the geodesic motion predicted by General Relativity. LARES can also be used to set limits on other theories of fundamental physics and for a number of measurements in space geodesy and geodynamics. LARES-type satellites, placed in different orbits, could be used for a number of other tests of gravitational and fundamental physics. For example, tests of Brane-World theories, further limits on some low-energy consequences of string theory, possibly related to dark energy and quintessence, Yukawa-type deviations from the standard inverse square law for gravity and further measurements of the so-called Post-Newtonian parameters, testing General Relativity against alternative gravitational theories, could be carried out.

## References

1. Misner, C. W., Thorne, K. S. & Wheeler, J. A., *Gravitation* (Freeman, San Francisco, 1973).
2. Misner, Hawking, S. W. & Ellis, G. F. R., *The Large Scale Structure of Space-Time* (Cambridge Univ. Press, 1975).
3. Ciufolini, I. & Wheeler, J. A., *Gravitation and Inertia* (Princeton Univ. Press, Princeton, New Jersey, 1995).

4. Will, C. M., *Theory and Experiment in Gravitational Physics* (Cambridge Univ. Press, Cambridge, UK, ed. 2, 1993).
5. Turyshev, S. G., Experimental Tests of General Relativity: Recent Progress and Future Directions, *Uspekhi Fizicheskikh Nauk* **197**(1), 3-34 (2009) [PhysicsUspekhi **52**(1), 3-34 (2009)].
6. Riess, A. et al, Observational evidence from supernovae for an accelerating universe and a cosmological constant, *Astron.* J. **116**, 1009-1038 (1998)
7. Perlmutter, S. et al., Measurements of V and L from 42 high-redshift supernovae, *Astrophys.* J. **517**, 565-586 (1999).
8. Perlmutter S., Supernovae, dark energy, and the accelerating universe, *Phys. Today* **56**, 3-59 (2003).
9. Caldwell, R. R., Dark energy, *Phys. World* **17**, 37-42 (2004).
10. Bertone, G., Hooper, D., Silk, J., Particle dark matter: evidence, candidates and constraints, *Physics Reports* **405**, 279-390 (2005).
11. Dvali, G., Gabadadze, G. & Porrati, M. 4D gravity on a brane in 5D Minkowskispace, *Phys. Lett. B* **485**, 208-214 (2000).
12. Carroll, S. et al, Is cosmic speed-up due to new gravitational physics?, *Phys. Rev. D* **70**, 083509 (2004).
13. Reinabelle Reyes, R., et al., Confirmation of general relativity on large scales from weak lensing and galaxy velocities, *Nature(Letters)* **464**, 256-258 (2010).
14. Hartle, J. B., *An Introduction to Einstein's General Relativity* (Addison Wesley, San Francisco, 2003).
15. Rindler, W., *Relativity: Special, General, and Cosmological* (Oxford University Press, Oxford, 2001).
16. Ehlers, J., *Survey of General Relativity Theory, in: Relativity, Astrophysics and Cosmology*, W. Israel Ed., 1-125 (Reidel Publ., 1973).
17. Ehlers, J., Geroch, R., Equation of motion of small bodies in relativity, *Annals of Phys.* **309**, 232-236 (2004).
18. Geroch, R. & Jang, P. S., Motion of a body in general relativity, *J. Math. Phys.* **16**, 65 (1975).
19. In General Relativity the Lorentzian metric describes the curved spacetime geometry, where light speed is the limit for information transfer. The Einstein tensor locally describes mass-energy in the Einstein equations, and the dominant energy condition ensures its overall positivity.
20. Ciufolini, I., Paolozzi, A., Pavlis, E. C., Ries, J. C., Koenig, R. & Matzner R., The LARES Space Experiment: LARES Orbit, Error Analysis and Satellite Structure, in: *General Relativity and John Archibald Wheeler*,Ciufolini I. & Matzner R. Eds., 467-492 (Springer, 2010).
21. Pearlman, M. R., Degnan J. J. & Bosworth J., The International Laser Ranging Service, *Adv. Space Res.* **30**, 135-143 (2002), DOI:10.1016/S0273-1177(02)00277-6.
22. Bencze, W.J., et al., On-orbit performance of the Gravity Probe B drag-freetranslation control system, *AAS* **06**, 083 (2006)
23. Bencze, LAGEOS scientific results, *J. Geophys. Res. B* **90**, 9215-9438 (1985).
24. Tapley, B. D., The GRACE mission: status and performance assessment, *Eos. Trans. AGU* **83**(47) Fall Meet. Suppl., Abstract G12B-01 (2002).
25. Reigber, Ch., Flechtner, F., Koenig, R., Meyer, U., Neumayer, K. et al., GRACEorbit and gravity field recovery at GFZ Potsdam-first experiences and perspectives, *Eos. Trans. AGU* **83**(47) Fall Meet. Suppl., Abstract G12B-03 (2002).
26. Pavlis, D. E. et al., GEODYN operations manuals, (Contractor Report, Raytheon, ITSS, Landover MD.1998).

27. Martin, C. F. & Rubincam, D. P., Effects of Earth albedo on the LAGEOS I satellite, *J. Geophys. Res. B* **101**, 3215-3226 (1996).
28. Rubincam, D. P. On the secular decrease in the semimajor axis of Lageos's orbit, *Cel Mech. and Dyn. Astr.* **26**, 361-382 (1982).
29. Rubincam, D. P, Yarkovsky thermal drag on LAGEOS, *J. Geophys. Res. B* **93**, 13805-13810 (1988).
30. Métris, G., Vokroulicky, D., Ries, J. C. & Eanes R. J., Nongravitational effects and the LAGEOS eccentricity excitations, *J. Geophys. Res.*, **102**, 2711-2729
31. Ciufolini, I. Dragging of Inertial Frames, *Nature*, **449**, 41-47 (2007).
32. Ciufolini I. & Pavlis, E. A Confirmation of the General Relativistic Prediction of the Lense-Thirring Effect, *Nature(Letters)*, **431**, 431, 958-960 (2004).
33. Ciufolini, I., Pavlis, E. C., Ries, J. C., Koenig, R., Sindoni, G., Paolozzi, A., and Neumayer, H. "Test of Gravitomagnetism with the LAGEOS and GRACE Satellites", in: *General Relativity and John Archibald Wheeler*, Ciufolini I. & Matzner R. Eds., 371-434 (Springer, 2010).
34. Smith,T. L., Erickcek,A. L., Caldwell, R. R. & Kamionkowski, M. The effects of Chern-Simons gravity on bodies orbiting the earth, *Phys. Rev. D*, **77**, 024015, 2008.

arXiv:1309.1699v1 [gr-qc] 5 Sep 2013

# Fundamental Physics and General Relativity with the LARES and LAGEOS satellites

Ignazio Ciufolini*[1,2], Antonio Paolozzi[3], Rolf Koenig[4], Erricos C. Pavlis[5], John Ries[6], Richard Matzner[7], Vahe Gurzadyan[8], Roger Penrose[9], Giampiero Sindoni[3] and Claudio Paris[3]

[1] *Dip. Ingegneria dell'Innovazione, Università del Salento, Lecce, Italy*
[2] *Centro Fermi, Rome, Italy*
[3] *Scuola di Ingegneria Aerospaziale, Sapienza Università di Roma, Italy*
[4] *Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences, Potsdam, Germany*
[5] *Goddard Earth Science and Technology Center (GEST), University of Maryland, Baltimore County, USA*
[6] *Center for Space Research, University of Texas at Austin, USA*
[7] *Center for Relativity, University of Texas at Austin, USA*

September 9, 2013

**Abstract**

Current observations of the universe have strengthened the interest to further test General Relativity and other theories of fundamental physics. After an introduction to the phenomenon of frame-dragging predicted by Einstein's theory of General Relativity, with fundamental astrophysical applications to rotating black holes, we describe the past measurements of frame-dragging obtained by the LAGEOS satellites and by the dedicated Gravity Probe B space mission. We also discuss a test of String Theories of Chern-Simons type that has been carried out using the results of the LAGEOS satellites. We then describe the LARES space experiment. LARES was successfully launched in

*ignazio.ciufolini@unile.it

February 2012 to improve the accuracy of the tests of frame-dragging, it can also improve the test of String Theories. We present the results of the first few months of observations of LARES, its orbital analyses show that it has the best agreement of any other satellite with the test-particle motion predicted by General Relativity. We finally briefly report the accurate studies and the extensive simulations of the LARES space experiment, confirming an accuracy of a few percent in the forthcoming measurement of frame-dragging.

# 1 Introduction

The current study of the universe and nature has on the one hand allowed us to understand some of the basic laws governing the infinitely small, down to spatial distances of the order of $10^{-16}$ cm, or less, corresponding to quarks, and on the other hand to observe and in part understand the evolution of the universe up to spatial distances of the order of $10^{28}$ cm corresponding to the position of some quasars and near the so-called big-bang. With regard to the time scale, the current study of the universe ranges from a few instants after the big-bang up to the present time, approximately 14 billion years later. The laws of physics cover the four fundamental interactions: gravitational, electromagnetic, weak and strong. The last three are encompassed in the Standard Model theory of gauge symmetries. Theories, not yet experimentally verified, such as String and Brane-World theories, try to unify gravitation with the other three interactions and to unify the two great physical theories of General Relativity and Quantum Mechanics. The goal is the unification of the four interactions of nature in a theory that can be experimentally tested, and that can also address one of the biggest mysteries and riddles of science, the composition of most of the universe in which we live, that is the nature of dark energy and dark matter. Indeed, the discovery of the accelerated universe [1, 2] is one of the outstanding events in science today and dark energy, or "quintessence", is regarded as a new exotic physical substance that is accelerating the expansion of the universe. Dark energy, together with dark matter, should constitute approximately 95% of the mass-energy of the universe in an unexplained form [3, 4, 5].

The evolution of the universe and the gravitational interaction are currently described by Einstein's gravitational theory of General Relativity [6]. General Relativity is a triumph of classical thought, created by Einstein to

satisfy the competing requirements of the Equivalence Principle (local inertial physics can show no evidence of gravity), and the large scale effects of gravity. Einstein's gravitational theory succeeded by postulating that gravitation is the curvature of spacetime and it is a fundamental component for understanding the universe that we observe. During the past century General Relativity achieved an experimental triumph [7, 8, 9]. On the one hand, a number of key predictions of Einstein's gravitational theory have been experimentally confirmed with impressive accuracy. On the other hand, General Relativity today has practical applications in space research, geodesy, astronomy and navigation in the Solar System, from the Global Navigation Satellite Systems (GNSS) to the techniques of Very Long Baseline Interferometry (VLBI) and Satellite Laser Ranging (SLR), and is a basic ingredient for understanding astrophysical and cosmological observations such as the expanding universe and the dynamics of binary systems of neutron stars.

Despite being a well verified description of gravity, General Relativity has encountered somewhat unexpected developments in observational cosmology and is affected by some theoretical problems. Indeed, the study of distant supernovae in 1998 led to a discovery that they accelerate away from us. Since then, what is now referred to as dark energy is at the center of attention of many theoreticians. Observational data currently support its interpretation as the cosmological constant introduced by Einstein. However its current value, comparable with the critical density, needs to be reconciled with the expectations of quantum field theory or analogous fundamental theory (e.g., [10]). Combining gravity with quantum field theory may be expected to reveal the nature of dark energy and hence resolve the mystery of its value, and whether it might be related to dark matter.

Among its theoretical problems, General Relativity predicts the occurrence of spacetime singularities [11], events in which every known physical theory ceases to be valid, the spacetime curvature diverges and time ends. Furthermore, General Relativity is a classical theory that does not include Quantum Mechanics and no one has succeeded in a quantized version of General Relativity, though this is a serious ongoing effort, with both Loop Quantum Gravity and String Theory approaches. Even though a breakdown of General Relativity should occur at the quantum level, some viable modifications of Einstein's theory already give different predictions at the classical level and might explain the riddle of the dark energy. Modifications of General Relativity on cosmological scales, for instance the so called f(R) theories (with higher order curvature terms in the action), have been proposed to ex-

plain the acceleration of the universe without dark energy [12]. In summary, every aspect of Einstein's gravitational theory should be directly tested and the accuracy of the present measurements of General Relativity and of the foundations of gravitational theories should be further improved.

## 2 Frame-dragging

The observational tests of gravitational physics divide into purely solar system measurements of various effects, binary pulsars observations, and intermediate- and long- range cosmological observations via gravitational radiation. Purely solar system measurements include redshift and clock measurements, light deflection, time-delay of electromagnetic waves, Lunar Laser Ranging (LLR), geodetic precession and frame-dragging measurements.

Today, among the main challenges in experimental gravitation, we have the direct detection of gravitational waves, the improved measurement of the Post-Newtonian parameters testing General Relativity versus alternative gravitational theories, improved tests of the Equivalence Principle and the accurate measurement of frame-dragging and gravitomagnetism. Frame-dragging, or dragging of inertial frames, and gravitomagnetism and are produced by mass-energy currents, e.g., by the angular momentum of a body, in the same way as magnetism is generated by electric-currents in electrodynamics [13].

The origin of inertia has intrigued scientists and philosophers for centuries and the inertial frames are at the foundations of physics and General relativity. What determines an inertial frame? In the Newtonian gravitational theory an inertial frame has an absolute existence, uninfluenced by the matter in the Universe. In Einstein's gravitational theory the *local* inertial frames have a key role [6, 14, 9]. The strong equivalence principle, at the foundations of General Relativity, states that the gravitational field is locally 'unobservable' in the freely falling frames and thus, in these local inertial frames, all the laws of physics are the laws of Special Relativity. However, the local inertial frames are determined, influenced and dragged by the distribution and flow of mass-energy in the Universe. The axes of these local, inertial frames are determined by free-falling torque-free test-gyroscopes, i.e., sufficiently small and accurate spinning tops. Therefore, these gyroscopes are dragged by the motion and rotation of nearby matter [6, 14, 9], i.e., their orientation changes with respect to the distant stars: this is the 'dragging of

inertial frames' or 'frame-dragging', as Einstein named it in a letter to Ernst Mach [15]. Frame-dragging represents in Einstein's theory the remnant of the ideas of Mach on the origin of inertia. Mach thought that centrifugal and inertial forces are due to rotations and accelerations with respect to all the masses in the Universe and this is known as Mach's principle [9].

In General Relativity, a torque-free spinning gyroscope defines an axis non-rotating relative to the local inertial frames, however, the orbital plane of a test particle is also a kind of gyroscope. Frame-dragging also has an intriguing influence on the flow of time and on electromagnetic waves propagating around a spinning body. Indeed, synchronization of clocks all around a closed path near a spinning body is not possible [16, 17] in any rigid frame not rotating relative to the 'fixed stars', because light corotating around a spinning body would take less time to return to a starting point (fixed relative to the 'distant stars') than would light rotating in the opposite direction [16, 17, 18, 19, 20]. Since frame-dragging affects clocks, light, gyroscopes [46, 47] (e.g., the gyroscopes of GP-B space experiment) and orbiting particles [21] (see sections 4 and 5 on the LAGEOS satellites and on the LARES space experiment), it also affects matter orbiting and falling on a spinning body. Indeed, an explanation of the constant orientation of the spectacular jets from active galactic nuclei and quasars, emitted in the same direction during a time that may reach millions of years, is based on frame-dragging of the accretion disk due to a super-massive spinning black hole [22, 23] acting as a gyroscope.

The precession $\dot{\mathbf{\Omega}}_{Spin}$ of the spin axis of a test-gyroscope by the angular momentum $\mathbf{J}$ of the central body is: $\dot{\mathbf{\Omega}}_{Spin} = \frac{3G((\mathbf{J}\cdot\hat{r})\hat{r}-\mathbf{J})}{c^2 r^3}$, where $\hat{r}$ is the position unit-vector of the test-gyroscope and r is its radial distance from the central body. Similarly to a small gyroscope, the orbital plane of a planet, moon or satellite is a huge gyroscope that feels general relativistic effects. Indeed, frame-dragging produces a change of the orbital angular momentum vector of a test-particle, i.e., the Lense-Thirring effect, that is, the precession of the nodes of a satellite, i.e., the rate of change of its nodal longitude: $\dot{\mathbf{\Omega}}_{Lense-Thirring} = \frac{2G\mathbf{J}}{c^2 a^3 (1-e^2)^{3/2}}$, where $\mathbf{\Omega}$ is the longitude of the nodal line of the satellite (the intersection of the satellite orbital plane with the equatorial plane of the central body), $\mathbf{J}$ is the angular momentum of the central body, $a$ the semi-major axis of the orbiting test-particle, $e$ its orbital eccentricity, $G$ the gravitational constant and $c$ the speed of light. A similar formula also holds for the rate of change of the longitude of the pericentre of

a test–particle, that is, of the so-called Runge-Lenz vector [21, 9].

Frame-dragging phenomena, which are due to mass currents and mass rotation, may be usefully described by a formal analogy of General Relativity, in a weak gravitational field and for slow motion, with electrodynamics (see Fig. 1) [23, 9] and have been called gravitomagnetism. Whereas an electric charge generates an electric field and a current of electric charge generates a magnetic field, in Newtonian gravitational theory the mass of a body generates a gravitational field but a current of mass, for example the rotation of a body, would not generate any additional gravitational field. On the other hand, Einstein's gravitational theory predicts that a current of mass would generate a gravitomagnetic field that would exert a force on surrounding bodies and would change the spacetime structure by generating additional curvature [24]. Furthermore, in General Relativity a current of mass in a loop (that is, a gyroscope) has a behaviour formally similar to that of a magnetic dipole in electrodynamics, which is made of an electric current in a loop. Then, the gravitomagnetic field generates frame-dragging of a gyroscope, in a similar way to the magnetic field producing the change of the orientation of a magnetic needle (magnetic dipole). In General Relativity, the gravitomagnetic field, $\mathbf{H}$, due to the angular momentum $\mathbf{J}$ of a central body is, in the weak-field and slow-motion approximation: $\mathbf{H} = \nabla \times \mathbf{h} \cong 2\,G \left[ \frac{\mathbf{J} - 3(\mathbf{J} \cdot \hat{x})\,\hat{x}}{c^3 r^3} \right]$, where $r$ is the radial distance from the central body, $\hat{x}$ is the position unit-vector and $\mathbf{h}$ is the so-called 'gravitomagnetic vector potential' (equal to the non-diagonal, space and time, part of the metric), see Fig. 1.

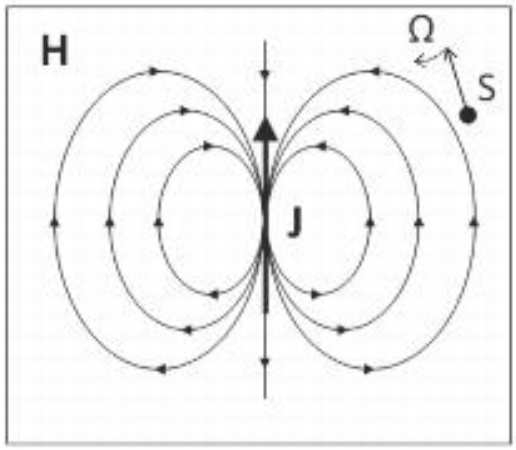



Figure 1: Frame-dragging and the gravitomagnetic analogy of General Relativity with electrodynamics. In General Relativity, freely falling test-gyroscopes define axes *fixed* relative to the local inertial frames, where the equivalence principle holds, that is, where the gravitational field is locally 'unobservable'; if we would rotate with respect to these gyroscope, we would then feel centrifugal forces, even though we may not rotate at all with respect to the 'distant stars', contrary to our everyday intuition. Indeed, a gyroscope is dragged by spinning masses, that is, its orientation changes with respect to the 'distant stars'. In this figure we show the gravitomagnetic field [47] $\mathbf{H}$ generated by the spin $\mathbf{J}$ of a central body and frame dragging $\dot{\mathbf{\Omega}}$ of a test gyroscope $\mathbf{S}$.

Since frame-dragging is due to the additional spacetime curvature produced by the rotation of a mass, to precisely characterize these phenomena, it has been proposed to use spacetime curvature invariants built using the Riemann curvature tensor (see: [25, 26] and section 6.11 of [9]). For discussions on the meaning of frame-dragging and gravitomagnetism, see: [27, 28, 29, 30, 31, 32, 33, 25, 26] and section 6.11 of [9].

# 3 String Theories and the LAGEOS and LARES Satellites

Among the extensions of General Relativity, the Chern-Simons gravity [34] with the Pontryagin density coupled scalar field in the Einstein-Hilbert ac-

tion, has attracted particular attention, since Chern-Simons gravitational term also emerges from String theories and Loop Quantum Gravity (see e.g. [36, 37] and references therein). The Pontryagin scalar is: ${}^*R^{\alpha\beta\mu\nu}\, R_{\alpha\beta\mu\nu}$ , that is a pseudoinvariant built "multiplying" the Riemann tensor $R^{\alpha\beta\mu\nu}$ with its dual ${}^*R^{\alpha\beta\mu\nu} \equiv \frac{1}{2}\varepsilon^{\alpha\beta\sigma\rho} R_{\sigma\rho}{}^{\mu\nu}$, where $\varepsilon^{\alpha\beta\sigma\rho}$ is the Levi Civita pseudotensor [35]. Due to the general character of Chern-Simons terms, the coupling constants can be even informative about the electroweak and even Planck scales [37], therefore any experimental constraint and even a null one can be of particular interest.

Concerning applications, Chern-Simons gravity has been involved to the interpretation of such basic cosmological and astrophysical problems as the dark energy, inflation, the evolution of binary neutron stars, gravitational wave emission by binary back holes and even the accretion powered energetic activity in the galactic nuclei and quasars [37, 38, 39, 40, 41]. Astrophysical observations, however, still do not allow to obtain constraints on Chern-Simons terms, e.g., as it is in the case even for the binary pulsar J0737-3039 [4], and the frame-dragging measurements near Earth are currently the only reasonable means to constraint the theory.

In 2008, Smith et al. [42] showed that String Theories of the type of Chern-Simons gravity predict an additional drift of the nodes of a satellite orbiting a spinning body and of a gyroscope spin axis. Then, using the frame-dragging measurement obtained with the LAGEOS satellites, they constrained the coupling constant of Chern-Simons theory (which may also be related to dark energy and quintessence, and to more fundamental parameters, such as related to a quintessence field). In particular, they set the lower limit to the Chern-Simons mass: $|m_{CS}| \gtrsim 0.001 km^{-1}$. See Fig. 2. Higher accuracy measurements by the LARES satellite will enable to improve that limit [42, 43].

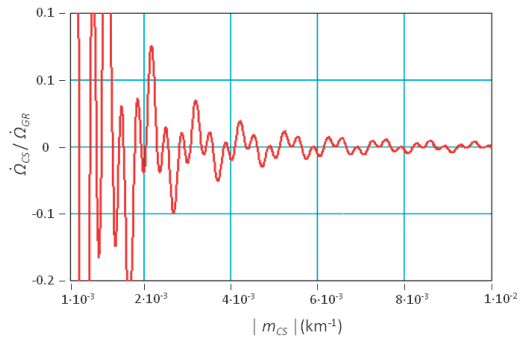



Figure 2: The ratio of the nodal rate of the LAGEOS satellites predicted by Chern-Simons gravity over that predicted by General Relativity implying the lower limit on the Chern-Simons mass: $|m_{CS}| \gtrsim 0.001 km^{-1}$ (adapted from [42]).

# 4 Tests of Frame-Dragging with the LAGEOS satellites and Gravity Probe-B

Since 1896 researchers, influenced by the ideas of Ernst Mach, tried to measure the frame-dragging effects generated by the rotation of the Earth on torsion balances [44] and gyroscopes [45]. In 1916, on the basis of General Relativity, de Sitter derived the Mercury perihelion precession due to the Sun angular momentum and, in 1918, Lense and Thirring [21] gave a general weak-field description of the frame-dragging effect on the orbit of a test-particle around a spinning body, today known as Lense-Thirring effect (see section 4). In 1959 and 1960, an experiment to test the general relativistic drag of a gyroscope was suggested [46, 47]. On 20 April 2004, after more than 40 years of preparation, the Gravity Probe B spacecraft was finally launched in a polar orbit at an altitude of about 642 km. The Gravity Probe B mission [50] (see http://einstein.stanford.edu/) consisted of an Earth satellite carrying four gyroscopes and one telescope, pointing at the guide star IM Pegasi (HR8703), and was designed to measure the drifts predicted by General Relativity (frame-dragging and geodetic precession) of the four test-gyroscopes with respect to the distant 'fixed' stars. General Relativity predicts that the average frame-dragging precession of the four Gravity Probe Bs gyroscopes by the Earths spin is about 39 milliarcseconds

per year (that is 0.000011 degrees per year) about an axis contained in Gravity Probe B's polar orbital plane. On 14 April 2007, after about 18 months of data analysis, the first Gravity Probe B results were presented: the Gravity Probe B experiment was affected by large drifts of the gyroscopes' spin axes produced by classical torques on the gyroscopes. The Gravity Probe B team explained [48] (see also [49]) the large drifts of the gyroscopes as being due to electrostatic patches on the surface of rotors and housings, and estimated the unmodeled systematic errors to be of the order of 100 milliarcseconds per year, corresponding to an uncertainty of more than 250% of the frame-dragging effect by the Earth spin. In 2011, finally, the Gravity Probe B team claimed that by some modeling of the systematic errors they were able to reduce the uncertainty in the measurement of frame-dragging to 19 % [50]. Frame-dragging is extremely small for Solar System objects, so to measure its effect on the orbit of a satellite we need to measure the position of the satellite to extremely high accuracy. Laser-ranging is the most accurate technique for measuring distances to the Moon and to artificial satellites such as LAGEOS (LAser GEOdynamics Satellite) [51]. Ultrashort-duration laser pulses are emitted from lasers on Earth and then reflected back to the emitting laser-ranging stations by retro-reflectors on the Moon or on artificial satellites. By measuring the total round-trip travel time of a laser pulse we are today able to determine the instantaneous distance of a retro-reflector on the LAGEOS satellites with a precision of a few millimeters [52] and their nodal longitude with an uncertainty of a fraction of a milliarcsec per year [53, 54, 55]. In 1976, LAGEOS was launched by NASA and, in 1992, LAGEOS 2 was launched by the Italian Space Agency and NASA. They have altitudes of approximately 5,900 km and 5,800 km respectively. The LAGEOS satellites' orbits can be predicted, over a 15-day period, with an uncertainty of just a few centimeters [53, 54, 55]. The Lense-Thirring drag of the orbital planes of LAGEOS and LAGEOS 2 is [56, 57] approximately 31 milliarcseconds per year, corresponding at the LAGEOS altitude to approximately 1.9 m per year. Since using laser-ranging we can determine their orbits with an accuracy of a few centimeters, the Lense-Thirring effect can be measured very accurately on the LAGEOS satellites' orbits if all their orbital perturbations can be modeled well enough [56, 57, 53]. On the other hand, the LAGEOS satellites are very heavy spherical satellites with small cross-sectional areas, so atmospheric particles and photons can only slightly perturb their orbits and especially they can hardly change the orientation of their orbital planes [57, 53, 58, 59]. By far the main perturbation of their

orbital planes is due to the Earth's deviations from spherical symmetry and by far the main error in the measurement of frame-dragging using their orbits is due to the uncertainties in the Earth's even zonal spherical harmonics [60]. The Earth's gravitational field and its gravitational potential can be expanded in spherical harmonics and the even zonal harmonics are those harmonics of even degree and zero order. These spherical harmonics, denoted as $J_{2n}$, where $2n$ is their degree, are those deviations from spherical symmetry of the Earth's gravitational potential that are axially symmetric and that are also symmetric with respect to the Earth's equatorial plane: they produce large secular drifts of the nodes of the LAGEOS satellites. In particular, the flattening of the Earth's gravitational potential, corresponding to the second degree zonal harmonic $J_2$ describing the Earth's quadrupole moment, is by far the largest error source in the measurement of frame-dragging since it produces the largest secular perturbation of the node of LAGEOS [56, 61]. But thanks to the observations of the geodetic satellites, the Earth's shape and its gravitational field are extremely well known. For example, the flattening of the Earth's gravitational potential is today measured [62] with an uncertainty of only about one part in $10^7$ that is, however, still not enough to test frame-dragging. To eliminate the orbital uncertainties due to the errors in the Earth's gravity models, the use of both LAGEOS and LAGEOS2 was proposed [61]. However, it was not easy to confidently assess the accuracy of some earlier measurements [63] of the Lense-Thirring effect with the LAGEOS satellites, given the limiting factor of the uncertainty of the gravity models available in 1998. In March 2002, the problem of the uncertainties in the Earth's gravity field was overcome when the twin GRACE (Gravity Recovery And Climate Experiment) [64, 65] spacecraft of NASA were launched in a polar orbit at an altitude of approximately 400 km and about 200-250 km apart. The spacecraft range to each other using radar and they are tracked by the Global Positioning System (GPS) satellites. The GRACE satellites have greatly improved our knowledge of the Earth's gravitational field. Indeed, by using the two LAGEOS satellites and the GRACE Earth gravity models, the orbital uncertainties due to the modeling errors in the non-spherical Earth's gravitational field are only a few per cent of the Lense-Thirring effect [66, 67, 68]. The method to measure the Lense-Thirring effect is to use two *observables*, provided by the two nodes of the two LAGEOS satellites, for the two unknowns: Lense-Thirring effect and uncertainty in the Earth quadrupole moment $\delta J_2$ [61]. In 2004, nearly eleven years of laser-ranging data were analyzed. This analysis resulted in a measurement of

the Lense-Thirring effect with an accuracy [66, 13, 67, 68] of approximately 10%. The uncertainty in the largest Earth's even zonal harmonic, that is the quadrupole moment $J_2$, was eliminated by the use of the two LAGEOS satellites, see Fig. 3. However, the main remaining error source was due to the uncertainty in the Earth even zonal harmonics of degree strictly higher than two and especially to the even zonal harmonic of degree four, i.e., $J_4$. After 2004, other accurate Earth gravity models have been published using longer GRACE observations. The LAGEOS analyses have then been independently repeated with new models, over a longer period and by using three different orbital programs developed by NASA Goddard, the University of Texas at Austin [69], see Fig. 4, and the German GeoForschungsZentrum (GFZ) Potsdam [70], see Fig. 5. The recent frame-dragging measurements [67, 68, 70] by a team from the universities of Salento, Rome, Maryland, NASA Goddard, the University of Texas at Austin and the GFZ Potsdam, have confirmed the 2004 LAGEOS determination of the Lense-Thirring effect. No deviations from the predictions of General Relativity have been observed.

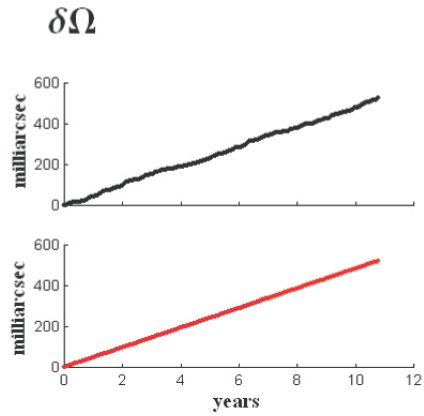



Figure 3: The 2004 measurement of frame-dragging using the LAGEOS and LAGEOS 2 satellites [66, 88]. The figure shows the *observed* orbital residuals of the nodal longitudes, $\delta\Omega$, of the LAGEOS satellites, combined in a suitable way to eliminate the uncertainty of the Earth's quadrupole moment. In black is the raw, observed, residual nodal longitude of the LAGEOS satellites after removal of six periodic signals. The best-fit line through these observed residuals has a slope of 47.9 mas $yr^{-1}$. In red is the theoretical Lense-Thirring prediction of Einstein's general relativity for the combination of the nodal longitudes of the LAGEOS satellites; its slope is 48.2 milliarcsec $yr^{-1}$ (adapted from [66]).

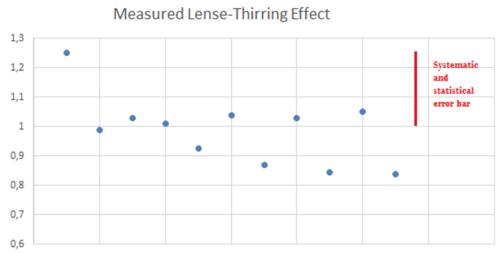



Figure 4: Independent 2008 measurement of frame-dragging using LAGEOS and LAGEOS 2 obtaned by CSR of the University of Texas at Austin using UTOPIA and the GRACE models: EIGEN-GRACE02S, GGM02S, EIGEN-CG03C, GIF22a, JEM04G, EIGEN-GL04C, JEM01-RL03B, GGM03S, ITG-GRACE03S and EIGEN-GL05C. The mean value of frame-dragging measured by Ries et al. using these models is 0.99 of the prediction of General Relativity. The total error budget of CSR-UT in the measurement of frame-dragging is about 12%.; see [69].

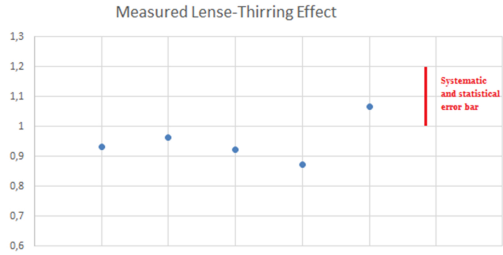



Figure 5: Independent 2012 measurement of frame-dragging using LAGEOS and LAGEOS 2 obtaned by GFZ Potsdam using EPOS-OC and the GRACE model: EIGEN-6C, EIGEN-6C (without considering trend and annual and semi-annual variations in the Earth gravitational field), EIGEN-6Sp.34, EIGEN-51C and EIGEN-GRACE03S. The mean value of frame-dragging measured by König et al. using these models is 0.95 of the prediction of General Relativity; see [70].

# 5 The LARES Space Experiment

In the test of frame-dragging using LAGEOS and LAGEOS 2, the main error source is due to the even zonal harmonic of degree four, $J_4$; such an error can

be as large as 10% of the Lense-Thirring effect [71]. Thus, to significantly increase the accuracy of the measurement of frame-dragging, one would need to eliminate that uncertainty by using an additional observable, i.e., by using a laser-ranged satellite in addition to LAGEOS and LAGEOS 2.

LARES (LAser RElativity Satellite) is a laser-ranged satellite of the Italian Space Agency (ASI), see Fig. 6. It was launched successfully on the 13th of February 2012 with the qualification flight of VEGA, the new launch vehicle of the European Space Agency (ESA), which was developed by ELV (Avio-ASI) [72, 73]. LARES, together with the LAGEOS and LAGEOS 2 satellites and the GRACE mission [64, 65], will provide an accurate test of Earth's frame-dragging with uncertainty of a few percent and other tests of fundamental physics [71, 68, 74]. The Lense-Thirring drag of the orbital planes of the LARES is approximately 118 milliarcseconds per year corresponding, at the LARES altitude, to approximately 4.5 m/yr.

The LARES orbital elements are as follows: the semi-major axis is 7820 km, orbital eccentricity 0.0007, and orbital inclination $69.5^o$. It is currently successfully tracked by the global International Laser Ranging Service (ILRS) station network [76]. LARES has the highest mean density of any known object orbiting in the Solar System. It is spherical and covered with 92 retro-reflectors, and it has a radius of 18.2 cm. It is made of a tungsten alloy, with a total mass of 386.8 kg, resulting in a ratio of cross-sectional area to mass that is about 2.6 times smaller than that of the two LAGEOS satellites [73]. Before LARES, the LAGEOS satellites had the smallest ratio of cross-sectional area to mass of any artificial satellite, such a ratio is critical to reduce the size of the non-gravitational perturbations. Indeed, the extremely small cross-sectional area to mass ratio of LARES, i.e. 0.00027 $m^2/kg$, and its special structure, a single piece solid sphere with high thermal conductivity, ensure that the unmodeled non-gravitational orbital perturbations are smaller than for any other satellite, in spite of its lower altitude compared to LAGEOS. This behavior has been confirmed experimentally using the first few months of laser ranging observations [72].

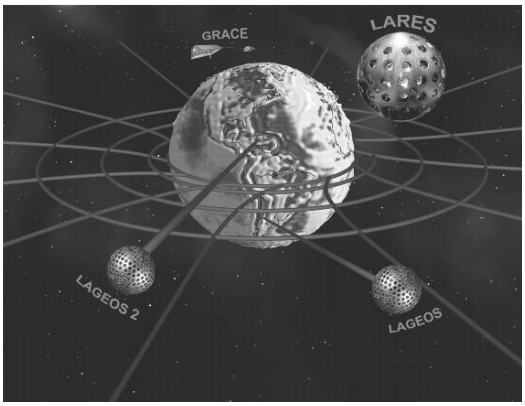



Figure 6: Artistic view of the LARES space experiment with the satellites: LARES, LAGEOS, LAGEOS 2 and GRACE. The radial twisted curves are an artistic representation of the spacetime twist owed to frame-dragging by the Earth rotation. The Earth is displayed using the gravitational field determination EIGEN-GRACE02S obtained with GRACE.

## 5.1 First results of LARES orbital analysis

At the very foundation of General Relativity is the geodesic motion of a small, structureless test-particle. Depending on the physical context, a star, planet or satellite can behave very nearly like a test-particle, so geodesic motion is used to calculate the advance of the perihelion of a planet's orbit, the dynamics of a binary pulsar system and of an Earth-orbiting satellite (a timelike geodesic path in spacetime's Lorentzian geometry is one that locally maximizes proper time, in analogy with the length-minimizing property of Euclidean straight lines). Verifying geodesic motion is then a test of paramount importance to General Relativity and other theories of fundamental physics.

General Relativity explains the gravitational interaction as the curvature of spacetime generated by mass-energy and mass-energy currents via the Einstein field equations [6, 77, 9]. For example, the gravitational attraction of Earth on its Moon and artificial satellites is explained by General Relativity via the spacetime curvature generated by the Earth's mass. The motion of any test body within the gravitational field of another massive body, e.g., the motion of a 'small' satellite around the Earth, is simply determined by a geodesic of spacetime with curvature generated by the massive body. Moon

and artificial Earth satellites follow approximately geodesics of the spacetime with deviations from an ideal geodesic path due their finite size and to the non-gravitational forces acting on them. Thus, geodesic motion is at the foundation of General Relativity and of any other theory where the gravitational interaction is described by spacetime curvature dynamically generated by mass-energy. Therefore, the creation of the best possible approximation for the free motion of a test-particle, a spacetime geodesic, is a profound goal for experiments dedicated to the study of the spacetime geometry in the vicinity of a body, yielding high-precision tests of General Relativity and constraints on alternative gravitational theories.

A fundamental issue regards the approximation to a geodesic that is provided by the motion of an actually extended body. In General Relativity [78, 79] the problem of an extended body is subtle, due not only to the non-linearity of the equations of motion, but also to the need to deal with the internal structure of the compact body, constructed of continuous media, where kinetic variables and thermodynamic potentials are involved. Further, there may be intrinsically non-local effects arising from the internal structure of the extended body, such as tidal influences. Moreover, there are problems concerning the approximations that need to be made in order to describe a given extended body as a test-particle moving along a geodesic. These problems are related to the fact that many of the common Newtonian gravitational concepts such as the 'center of mass', 'total mass' or 'size' of an extended material body do not have well-defined counterparts in General Relativity [80]. The Ehlers-Geroch theorem [81] (generalizing the result in [82]) attributes a geodesic to the trajectory of an extended body with a small enough own gravitational field, if for a Lorentzian metric the Einstein tensor satisfies the so-called dominant energy condition [77], this tensor being non-zero in some neighborhood of the geodesics and vanishing at its boundaries. This theorem, asserting that small massive bodies move on near-geodesics, thus achieves a rigorous bridge from General Relativity to space experiments with 'small' satellites which suggests a high level of suppression of non-gravitational and self-gravitational effects from the satellite's own small gravitational field. This enables us to consider the satellite's motion to be nearly geodesic and hence provides a genuine testing ground for General Relativity's effects.

Given the extreme weakness of the gravitational interaction with respect to the other interactions of nature, the space environment is the ideal laboratory to test gravitational and fundamental physics. However, in order to

test gravitational physics, a satellite must behave as nearly as possible as a test-particle and must be as little as possible affected by non-gravitational perturbations such as radiation pressure and atmospheric drag. In addition, its position must be determined with extreme accuracy.

The best realization of an orbiting test-particle is LARES. By measuring the total round-trip travel time of a laser pulse, it is possible to determine the instantaneous distance to the satellite with an accuracy of a few millimeters. However, in order to test gravitational physics, we not only need to measure the position of a body with extreme accuracy, but we also need it to behave like a test-particle. In space, a test-particle can be realized in two ways: a small drag-free satellite or a small spacecraft with high density and an extremely small area-to-mass ratio. In the case of the drag-free Gravity Probe-B satellite, a mean residual acceleration of about $40 \times 10^{-12}\, m/s^2$ was achieved [21]. For a passive satellite (with no drag-free system), the key characteristic that determines the level of attenuation of the non-gravitational perturbations is the density, reflected by the ratio between its cross-sectional area and its mass.

We processed the LARES laser ranging data based on the first seven 15-day arcs using the orbital analysis and data reduction systems UTOPIA of UT/CSR (Center for Space Research of The University of Texas at Austin), GEODYN II of NASA Goddard, and EPOS-OC of GFZ (Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences) [83]. In all cases, state-of-the art satellite orbital dynamical models were employed, including all the general relativistic post-Newtonian corrections, GRACE-based mean gravity field models [64, 65], modern models for the ocean and solid Earth tides, as well as solar radiation pressure, Earth albedo and atmospheric drag [84, 85, 58]. No 'thermal thrust' [86, 87] models were used. For the 105 days analyzed, GEODYN, UTOPIA and EPOS-OC independently determined that the residual along-track accelerations for LARES were only about $0.4 \times 10^{-12}\, m/s^2$, whereas for the two LAGEOS satellites, the acceleration residuals were 1-2 $\times 10^{-12}\, m/s^2$.

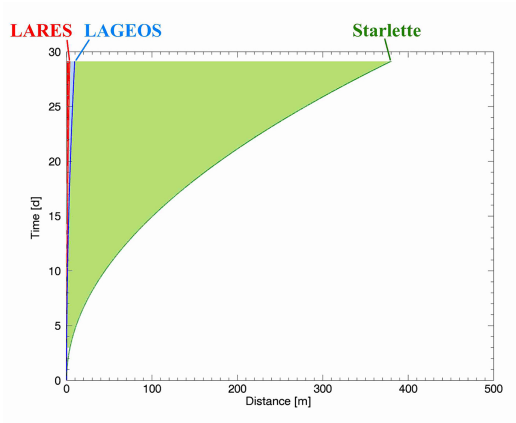



Figure 7: The red curve represents the change of distance between a 'test-particle' following a spacetime geodesic, represented here by the axis of ordinates in a frame co-moving with the test-particle, and a similar particle perturbed by the average unmodelled along-track acceleration of the magnitude observed on the LARES satellite of approximately $0.4 \times 10^{-12}\, m/s^2$. The blue and green curves represents the change of distance between a test-particle and a similar particle perturbed by an average along-track acceleration of the typical size of the unmodelled along-track acceleration observed on the LAGEOS satellites, of the order of $1 \times 10^{-12}\, m/s^2$, and respectively of STARLETTE, with a typical residual acceleration of the order of $40 \times 10^{-12}\, m/s^2$. The axis of ordinates may be thought of to represent a spacetime geodesic followed by LARES or LAGEOS after removing all the known and unmodelled non-gravitational perturbations (adapted from [72]).

This is particularly impressive given that LARES is far lower in the Earth's atmosphere than LAGEOS. The residual along-track accelerations of a satellite provide a measure of the level of suppression of its non-gravitational perturbations: atmospheric drag, solar and terrestrial radiation pressure and thermal-thrust effects. Atmospheric drag acts primarily along the satellite's velocity vector, while solar radiation pressure, terrestrial radiation pressure (the visible and infrared radiation from Earth) and thermal-thrust effects will all have some contribution along-track as well. We recall that the Yarkovsky effect on a spinning satellite is a thermal thrust resulting from the anisotropic temperature distribution over the satellite's surface caused by solar heating. A variation of this effect, due to the Earth's infrared radiation, is the Earth-Yarkovsky or Yarkovsky-Rubincam effect [86, 87].

The effects of the residual unmodelled along-track acceleration on the orbits of the laser ranged satellites: LARES, LAGEOS and STARLETTE (a CNES laser ranged satellite launched in 1975) are illustrated in fig. 7 where we plot the change in the distance from their 'ideal' orbit, caused by the unmodelled along-track accelerations [72]. The vertical axis may be thought of as representing an 'ideal' reference world line of LARES, LAGEOS and Starlette, 'ideal' in the sense that all of its orbital perturbations are known. Figure 7 shows the unmodelled deviations from geodesic motion for LARES, LAGEOS and Starlette (once the known non-gravitational perturbations are removed, to the extent permitted by our current models) due to the unmodelled along-track accelerations. In these figures, we show the effect of a typical residual unmodelled along-track acceleration of $1 \times 10^{-12}\, m/s^2$ for LAGEOS, $0.4 \times 10^{-12}\, m/s^2$ for LARES and $40 \times 10^{-12}\, m/s^2$ for Starlette. Since all the general relativistic post-Newtonian corrections were included in our orbital analyses, these figures show the level of agreement of the LARES and LAGEOS orbits with the geodesic motion predicted by General Relativity.

It must be stressed that a residual unmodelled out-of-plane acceleration, constant in direction, of the order of magnitude of the unmodelled along-track acceleration observed on LARES, will produce an extremely small secular variation of the longitude of its node, i.e., of its orbital angular momentum. For example, by considering an out-of-plane acceleration with amplitude of $0.4 \times 10^{-12}\, m/s^2$ constant in direction, its effect on the node of LARES would be many orders of magnitude smaller than the tiny secular drift of the node of LARES due to frame-dragging [30] of about 118 milliarcsec/y. Therefore, LARES, together with the LAGEOS satellites, and with the determination of Earth's gravitational field obtained by the GRACE mission, will be used to accurately measure the frame-dragging effect predicted by General Relativity, improving by about an order of magnitude the accuracy of previous frame-dragging measurements by the LAGEOS satellites [66, 67, 68].

In conclusion, LARES provides the best available test-particle in the Solar System for tests of gravitational physics and General Relativity, e.g., for the accurate measurement of frame-dragging, and after modelling its known non-gravitational perturbations, its orbit shows the best agreement of any satellite with the geodesic motion predicted by General Relativity.

## 5.2 Error analysis and Monte Carlo Simulations of the LARES experiment

A large number of papers have been published that analyze all the error sources, of both gravitational and non-gravitational origin, that can affect the LAGEOS and LARES experiments (see, e.g., [57, 53, 54, 61, 55, 88, 67, 71, 68, 74, 43]. The largest measurement uncertainties are due to the errors in the first two Earth even zonal harmonics, of degree 2 and 4, i.e., $\delta J_2$ and $\delta J_4$, but they are eliminated using three observables, i.e., the three nodes of the LARES, LAGEOS and LAGEOS 2 satellites, thus allowing a measurement of frame-dragging with an uncertainty of a few percent. Furthermore, the LARES inclination of $69.5^o$ minimizes the uncertainties due to the error in the Earth even zonal harmonics of degree higher than four, i.e., $\delta J_{2n}$ with $2n > 4$. This is the largest source of error in the measurement of frame-dragging using the LAGEOS. LAGEOS 2 and LARES satellites. The error in the LARES experiment due to each even zonal harmonic up to degree 70 was analyzed in detail in [71, 68]. The LARES error analyses have been recently confirmed by a number of Monte Carlo simulations [74].

In Fig. 8 we display the error in the LARES experiment due to each even zonal harmonic up to degree 70. In this figure, the largest errors due to the uncertainties in the first two even zonal harmonics, of degree 2 and 4, are not shown since they are eliminated in the measurement of frame-dragging using the 3 observables, i.e. the 3 nodes of LARES, LAGEOS and LAGEOS 2. Fig. 8 clearly displays that the error due to each even zonal harmonic of degree higher than 4 is considerably less than 1% and in particular that the error is substantially negligible for the even zonal harmonics of degree higher than 26.

The results of Fig. 8 are based on the calibrated uncertainties (i.e., including systematic errors) of the EIGEN-GRACE02S (GFZ, Potsdam, 2004) model (used in [66]). In Fig. 8 we also display the maximum percent errors due to each even zonal harmonic obtained by considering as uncertainty for each harmonic the difference between the value of that harmonic in the EIGEN-GRACE02S model minus its value in the GGM02S model (a model with comparable accuracy); this is a standard technique in space geodesy to estimate the reliability of the published uncertainties of a model; of course, in order to use this technique, one must use models of comparable accuracy, i.e., models that are indeed comparable, or use this technique only to assess the errors of the less accurate model.

Using EIGEN-GRACE02S and GGM02S (see [71]), the total error in the measurement of the Lense-Thirring effect due to the even zonal harmonics is respectively 1.4 % and 2.1 %. Even though the real error in the EIGEN-GRACE02S coefficients would probably be about two or three times larger than these published uncertainties, EIGEN-GRACE02S was just a preliminary 2004 determination of the Earth gravitational field and models much more accurate than EIGEN-GRACE02S, based on much longer GRACE observations. are today available. Indeed, these two models, EIGEN-GRACE02S and GGM02S have been obtained with a relatively small amount of observations of the GRACE spacecraft (launched in February 2002) and therefore a substantial factor of improvement over these two GRACE models has to be taken into account at the time of the LARES data analysis (between 2012 and 2018), thanks to longer GRACE observational periods and to other space geodesy missions too.

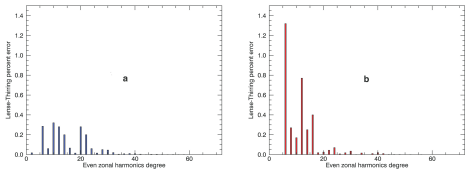



Figure 8: Percent error in the measurement of frame-dragging using LARES, LAGEOS and LAGEOS 2 as a function of the uncertainty due to each even zonal harmonic. The points in blue in panel $a$ are the errors obtained using the model EIGEN-GRACE02S and the points in red in panel $b$ are the errors obtained using as uncertainty of each coefficient the difference between the value of this coefficient in the two different models EIGEN-GRACE02S and GGM02S. The total error in the measurement of the Lense-Thirring effect using EIGEN-GRACE02S is 1.4 % and by using as uncertainties the differences between the coefficients of the two models is 3.4 %. However, at the time of the LARES data analysis a substantial improvement has to be taken into account with respect with these older 2004 models that were based on less than 365 days of observations of the GRACE spacecraft. Today, the GRACE determinations of the Earth gravitational field are already much more accurate than the two 2004 GRACE models used to derive the Earth gravitational field displayed in figure 8.

In regard to a detailed treatment of the other orbital perturbations that

affect the LARES experiment, tidal effects and non-gravitational perturbations, such as solar and albedo radiation pressure, thermal thrust and particle drag, we refer to [57, 53, 88, 67, 71]. In regard to the orbital perturbations on the LARES experiment due to the time dependent Earth's gravity field, we observe that the largest tidal signals are due to the zonal tides with $l = 2$ and $m = 0$, due to the Moon node, and to the $K_1$ tide with $l = 2$ and $m = 1$ (tesseral tide). However, the error due to the medium and long period zonal tides ($l = 2$ and $m = 0$) will be eliminated, together with the static $J_2$ error, using the combination of the three nodes (also the uncertainties in the time-dependent secular variations $\dot{J}_2$, $\dot{J}_4$ will be cancelled using this combination of three observables). Furthermore, the tesseral tide $K_1$ will be fitted for over a period equal to the LARES nodal period (see [53] and chapter 5 of [55]) and this tide would then introduce a small uncertainty in our combination. In regard to the non-gravitational orbital perturbations, we simply observe here that the LAGEOS satellites and especially the LARES satellite are extremely dense spherical satellites with very small cross-sectional-to-mass ratio in order to reduce their non-gravitational perturbations [57]. In particular, in the previous section 5.1 we have shown that the unmodelled perturbations of the LARES orbit, in spite of its lower orbit, are smaller than on the LAGEOS satellites owed to the much smaller cross-sectional-to-mass ratio of LARES and to its special structure. We finally point out that the neutral and charged particle drag on the LARES node is a negligible effect. That is owed to the almost circular orbit of LARES, i.e., its orbital eccentricity is $e \cong 0.0007$, and to the LARES special structure. Indeed, even assuming that the exosphere would be co-rotating with the Earth at any satellite altitude, in the case of zero orbital eccentricity, $e = 0$, the total nodal shift of the satellite would be zero, as calculated in [57]. Indeed, the nodal rate of a satellite due to particle drag is a function of $sin\ \nu \cdot cos\ \nu$ (where $\nu$ is the true anomaly) and the total nodal shift is then zero over one orbit. In the case of a very small orbital eccentricity, the total nodal shift would be proportional to the eccentricity and thus for LARES it would be a very small effect [57] owed also to its very small cross-sectional-to-mass ratio.

A number of Monte Carlo simulations have recently confirmed the previous detailed and extensive error analyses of the LARES experiment [74], i.e., the potentiality of the LARES experiment to achieve a measurement of frame-dragging with an uncertainty of a few percent only. These simulations have confirmed that the three observables provided by the three nodes of the LARES, LAGEOS and LAGEOS 2 satellites, together with the latest Earth

gravitational field determinations from the GRACE space mission, will allow us to improve significantly the previous measurements of the phenomenon of frame-dragging predicted by General Relativity, by eliminating the uncertainties in the value of the first two even zonal harmonics of the Earth potential $\delta J_2$ and $\delta J_4$.

The 100 simulations were designed to reproduce as closely as possible the real experiment to measure frame-dragging using LARES, LAGEOS, LAGEOS-2 and GRACE. We considered a number of physical parameters whose uncertainties have a critical impact on the accuracy of the measurement of the frame-dragging effect using LARES, LAGEOS and LAGEOS-2. Together with the values of these critical parameters, determined either by the GRACE space mission (in the case of the Earth gravitational field parameters) or by previous extensive orbital analyses (in the case of the radiation pressure parameters of the satellites), we consider their realistic uncertainty estimated by also taking into account the systematic errors. Then, using EPOS-OC, we simulated (100 times) the orbits of the LARES, LAGEOS and LAGEOS 2 satellites by randomly generating values of the GM (mass) of Earth, of its five largest even zonal harmonics, $J_2$, $J_4$, $J_6$, $J_8$ and $J_{10}$, of the secular rate of change of the two largest even zonal harmonics $\dot{J}_2$ and $\dot{J}_4$ and of the solar radiation coefficients of LARES, LAGEOS and LAGEOS 2. The frame-dragging effect was always kept equal to its General Relativity value. Finally, we carried out the analysis of their simulated laser-ranging observations.

The result of the 100 simulations of the LARES experiment was that the standard deviation of the measured simulated values of frame-dragging was equal to 1.4% of the frame-dragging effect predicted by General Relativity. Its mean value effect was equal to 100.24% of its general relativistic value. Thus, the Monte Carlo simulations confirmed an error budget of about 1% in the forthcoming measurement of frame-dragging using LARES, LAGEOS, LAGEOS 2 and GRACE.

# 6 Conclusions

Frame-dragging is an intriguing phenomenon predicted by General Relativity with fundamental astrophysical applications to rotating black holes. Past measurements of frame-dragging have been performed using the LAGEOS satellites and the dedicated Gravity Probe B space mission, respectively with

accuracies of about 10% and 19%. The LAGEOS tests of frame-dragging have been independently obtained by three teams: Universities of Salento, Sapienza and Maryland, University of Texas at Austin, and GFZ Potsdam, using three different orbital programs. The LAGEOS results were also used to constrain String Theories of Chern-Simons type. The LARES space experiment will improve the measurement of frame-dragging by one order of magnitude by also improving the test of String Theories. The orbital analyses of the first few months of observations of LARES have shown that the LARES orbit has the best agreement of any other satellite with the test-particle motion predicted by General Relativity. Accurate error analyses and extensive simulations have confirmed a total error of a few percent in the forthcoming measurement of frame-dragging using LARES, LAGEOS, LAGEOS 2 and GRACE.

# 7 Acknowledgements

The authors gratefully acknowledge the International Laser Ranging Service for providing high-quality laser ranging tracking of the LARES satellites. I. Ciufolini and A. Paolozzi gratefully acknowledge the support of the Italian Space Agency, grants I/043/08/0, I/016/07/0, I/043/08/1 and I/034/12/0. J.C. Ries the support of NASA Contract NNG06DA07C, and E.C. Pavlis and R.A. Matzner the support of NASA Grant NNX09AU86G.

# References

[1] A. Riess et al., Observational evidence from supernovae for an accelerating universe and a cosmological constant. *Astron. J.* **116**, 1009, (1998).

[2] S. Perlmutter et al., Measurements of $\Omega$ and $\Lambda$ from 42 High-Redshift Supernovae. *Astrophys. J.* **517**, 565, (1999).

[3] S. Perlmutter, Supernovae, Dark Energy, and the Accelerating Universe. *Phys. Today* **56**, 53, (2003).

[4] Robert R. Caldwell, "Dark Energy," *Physics World*, 37-42, (2004).

[5] Planck Collaboration, Planck 2013 results, submitted to *Astronomy and Astrophysics*, (2013).

[6] C.W. Misner, K.S. Thorne, and J.A. Wheeler, *Gravitation* Freeman, San Francisco, (1973).

[7] S. Turyshev, Experimental Tests of General Relativity: Recent Progress and Future Directions. *Physics-Uspekhi* **52**, 1, (2009).

[8] C. M. Will, *Theory and Experiment in Gravitational Physics 2nd edn* (Cambridge Univ. Press, Cambridge, UK, 1993).

[9] I. Ciufolini and J.A. Wheeler, *Gravitation and Inertia*,Princeton Univ. Press, (1995).

[10] M. Kamionkowski, in: *Visions of Discovery*, Cambridge Univ. Press, **247**, (2007).

[11] R. Penrose, Gravitational Collapse and Space-Time Singularities. *Phys. Rev. Lett.* **14**, 57, (1965).

[12] A. De Felice and S. Tsujikawa, f(R) Theories. *Living Rev. Relativ.* **13**, 3, (2010).

[13] I. Ciufolini, Dragging of Inertial Frames. *Nature* **449**, 41, (2007).

[14] S. Weinberg, *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity* (Wiley, New York, 1972).

[15] A. Einstein, Letter to Ernst Mach. Zurich, 25 June 1913, in ref. [6] p. 544.

[16] L. D. Landau, and E. M. Lifshitz, *The Classical Theory of Fields 3rd rev.*, English edn. (Pergamon, London, 1971).

[17] Ya. B. Zeldovich, and I. D. Novikov, *Relativistic Astrophysics Vol. I Stars and Relativity* (Univ. Chicago Press, Chicago, 1971).

[18] I. Ciufolini and F. Ricci, Time delay due to spin and gravitational lensing. *Class. and Quantum Grav.* **19**, 3863-3874 (2002).

[19] I. Ciufolini and F. Ricci, Time delay due to spin inside a rotating shell. *Class. and Quantum Grav.* **19**, 3875-3881 (2002).

[20] I. Ciufolini, F. Ricci, S. Kopekin and B. Mashhoon, On the Gravitomagnetic Time Delay, *Physics Letters A* **308**, 101-109 (2003).

[21] J. Lense and H. Thirring, Über den Einfluss der Eigenrotation der Zentralkörper auf die Bewegung der Planeten und Monde nach der Einsteinschen Gravitationstheorie. *Phys. Z.* **19** 156-163, (1918). See also English translation by B Mashhoon, F W Hehl, D S Theiss *Gen. Relativ. Gravit.* **16**, 711-750 (1984).

[22] J.M. Bardeen and J.A. Petterson, The Lense-Thirring Effect and Accretion Disks around Kerr Black Holes. *Astrophysical J.* **195**, L65-7 (1975).

[23] K.S. Thorne, R.H. Price and D.A. Macdonald, The Membrane Paradigm (Yale Univ. Press, NewHaven, 1986).

[24] R. P. Kerr, Gravitational field of a spinning mass as an example of algebraically special metrics, *Phys. Rev. Lett.* **11**, 237-238 (1963).

[25] I Ciufolini, Gravitomagnetism and status of the LAGEOS III experiment. *Class. Quantum Grav.* **11**, A73-A81 (1994).

[26] I. Ciufolini, Frame-Dragging, Gravitomagnetism, and Lunar Laser Ranging. *New Astronomy* **15**, 332-337 (2010).

[27] N. Ashby and B. Shahid-Saless, Geodetic Precession or Dragging of Inertial Frames? *Phys. Rev. D* **42**, 1118-22 (1990).

[28] R.F. O'Connell, A Note on Frame Dragging. *Class. Quant. Grav.* **22**, 3815-16 (2005).

[29] B.M. Barker and R.F. O'Connel, The gravitational interaction: Spin, rotation, and quantum effects: A review. *Gen. Rel. Grav.* **11**, 149-175 (1979).

[30] A.R. Khan and R.F. O'Connell, Gravitational analogue of magnetic force. *Nature* **261**, 480-481 (1976).

[31] T.W. Murphy Jr, K. Nordtvedt and S.G. Turyshev, Gravitomagnetic Influence on Gyroscopes and on the Lunar Orbit. *Phys. Rev. Lett.* **98**, 071102–1-4 (2007).

[32] S.M. Kopeikin, Comment on "Gravitomagnetic Influence on Gyroscopes and on the Lunar Orbit". *Phys. Rev. Lett.* **98**, 229001 (2007).

[33] T.W. Murphy Jr, K. Nordtvedt and S.G. Turyshev, Murphy, Nordtvedt, and Turyshev Reply. *Phys. Rev. Lett.* **98**, 229002 (2007).

[34] R. Jackiw and S.-Y. Pi, Chern-Simons modification of general relativity. *Phys.Rev. D* **68** 104012 (2003).

[35] A. Z. Petrov, New Methods in General Relativity, Nauka, Moscow. English edition: Einstein Spaces. Pergamon Press (1969).

[36] B. A. Campbell, M.J. Duncan, N. Kaloper and K. A. Olive, Gravitational dynamics with lorentz chern-simons termsm. *Nuclear Physics B* **351**, 778 (1991).

[37] S. Alexander and N. Yunes, Chern-Simons modified general relativity. *Phys. Rep.* **480** 1-55 (2009).

[38] K. Yagi, N. Yunes and T. Tanaka, Slowly Rotating Black Holes in Dynamical Chern-Simons Gravity: Deformation Quadratic in the Spin. *Phys. Rev.D* **86** 044037 (2012).

[39] S. Alexander, A. Marciano and D. Spergel, Chern-Simons Ination and Baryogenesis arXiv:1107.0318.

[40] T. Harko, Z. Kovács, F. S. N. Lobo,Thin accretion disk signatures in dynamical Chern-Simons modied gravity. *Class.Quant.Grav.* **27**:105010, (2010).

[41] K. Yagi, N. Yunes and T. Tanaka, Gravitational Waves from Quasicircular Black-Hole Binaries in Dynamical Chern-Simons Gravity. *Phys. Rev. Lett.* **109**, 251105 (2012).

[42] T.L. Smith, A. Erickcek, R. Caldwell and M. Kamionkowski, Effects of Chern-Simons gravity on bodies orbiting the Earth. *Phys. Rev.D* **77** 024015 (2008).

[43] V.G. Gurzadyan, I. Ciufolini, S. Sargsyan, G. Yegorian, S. Mirzoyan and A. Paolozzi. *EPL* **102** 60002-p1-p4 (2013).

[44] B. and I. Friedländer *Absolute und Relative Bewegung?* (Berlin: Simion-Verlag, 1896).

[45] A. Föppl, Über einen Kreiselversuch zur Messung der Umdrehungsgeschwindigkeit der Erde. Sitzb. Bayer. Akad. Wiss. **34**, 5–28 (1904), Phys. Z. **5**, 416; see also A. Föppl, Über Absolute und Relative Bewegung. *Sitzb. Bayer. Akad. Wiss.* **34**, 383–95 (1904).

[46] G.E. Pugh, Proposal for a Satellite Test of the Coriolis Prediction of General Relativity, *Weapons Systems Evaluation Group Research Memorandum N.* **11** (The Pentagon, Washington, 1959).

[47] L.I. Schiff, Motion of a Gyroscope According to Einstein's Theory of Gravitation. Proc. Nat. Acad. Sci. **46**, 871-82 (1960) and Possible New Test of General Relativity Theory. *Phys. Rev. Lett.* **4**, 215-7 (1960).

[48] D.K. Gill and S. Buchman, Evidence for Patch Effect Forces On the Gravity Probe B Gyroscopes. (Stanford Univ., Stanford, April 2007); poster at http://einstein.stanford.edu/content/aps_posters/EvidenceForPatchEffectForces.pdf.

[49] B.M. Barker and R.F. O'Connel, The gyroscope test of General Relativity. *Nature* **312**, 314 (1984).

[50] C.W. Everitt et al., 2011 Gravity Probe B: Final Results of a Space Experiment to Test General Relativity *Phys. Rev. Lett.* **106**, 22110 (2011).

[51] S.C. Cohen and P.J. Dunn, (Eds.), LAGEOS Scientific Results, *J. Geophys. Res.* **90** (B11), 9215 (1985).

[52] *13th International Workshop on Laser Ranging, Proceedings From the Science Session and Full Proceedings CD-ROM*, edited by R. Noomen, S. Klosko, C. Noll and M. Pearlman (NASA CP 2003-212248, NASA Goddard, Greenbelt, MD, 2003).

[53] B. Tapley, J.C. Ries, R.J. Eanes, and M.M. Watkins, *NASA-ASI Study on LAGEOS III*, CSR-UT publication n. CSR-89-3, Austin, Texas (1989), and I. Ciufolini et al., *ASI-NASA Study on LAGEOS III*, CNR, Rome, Italy (1989). See also: I. Ciufolini et al., *INFN study on LARES/WEBER-SAT* (2004).

[54] J.C. Ries , *Simulation of an experiment to measure the Lense-Thirring precession using a second LAGEOS satellite*, Ph. Dissertation, (Univ. of Texas, Austin, 1989).

[55] G.E. Peterson, *Estimation of the Lense-Thirring Precession Using Laser-Ranged Satellites*, Ph. Dissertation (Univ. of Texas, Austin, 1997).

[56] I. Ciufolini, Measurement of the Lense-Thirring drag on high-altitude laser-ranged artificial satellites. *Phys. Rev. Lett.* **56**, 278-281 (1986).

[57] I. Ciufolini, A comprehensive introduction to the Lageos gravitomagnetic experiment: from the importance of the gravitomagnetic field in physics to preliminary error analysis and error budget. *Int. J. Mod. Phys. A* **4**, 3083-3145 (1989).

[58] D.P. Rubincam, On the secular decrease in the semimajor axis of Lageos's orbit, *Celest. Mech.* **26**, 361-382 (1982).

[59] D.M. Lucchesi, Reassessment of the error modelling of non–gravitational perturbations on LAGEOS 2 and their impact in the Lense–Thirring determination. Part I, *Planet. Space Sci.* **49**, 447-463 (2001).

[60] W.M. Kaula, *Theory of Satellite Geodesy* (Blaisdell, Waltham, 1966).

[61] I. Ciufolini, On a new method to measure the gravitomagnetic field using two orbiting satellites. *Nuovo Cimento A* **109**, 1709-1720 (1996).

[62] G. Petit G and B. Luzum (eds.) 2010 IERS Conventions, Frankfurt am Main *Verlag des Bundesamts fr Kartographie und Geodaesie* 179 pp., ISBN 3-89888-989-6.

[63] I. Ciufolini, E.C. Pavlis, F. Chieppa, E. Fernandes-Vieira, and J. Perez-Mercader, Test of general relativity and measurement of the Lense-Thirring effect with two Earth satellites. *Science* **279**, 2100-2103 (1998).

[64] Ch. Reigber, F. Flechtner, R. Koenig, U. Meyer, K. Neumayer, R. Schmidt, P. Schwintzer and S. Zhu, GRACE Orbit and Gravity Field Recovery at GFZ Potsdam - First Experiences and Perspectives, *Eos. Trans. AGU*, **83**(47), Fall Meet. Suppl., Abstract G12B-03 (2002).

[65] B.D. Tapley, The GRACE Mission: Status and Performance Assessment, *Eos. Trans. AGU*, **83**(47), Fall Meet. Suppl., Abstract G12B-01 (2002).

[66] I. Ciufolini and E.C. Pavlis, A confirmation of the general relativistic prediction of the Lense-Thirring effect, *Nature* **431**, 958-960, (2004).

[67] I. Ciufolini, E.C. Pavlis, J. Ries, R. Koenig, G. Sindoni, A. Paolozzi and H. Newmayer, Gravitomagnetism and its Measurement with Laser Ranging to the LAGEOS satellites and GRACE Earth Gravity Models, in *John Archibald Wheleer and General Relativity* I. Ciufolini and R. Matzner eds., 371-434 (Springer Verlag, 2010).

[68] I. Ciufolini, A. Paolozzi, E.C. Pavlis, J. Ries, R. Koenig, R. Matzner, G. Sindoni and H. Neumayer, Testing Gravitational Physics with Satellite Laser Ranging *The European Physical Journal Plus* **126** 72 (2011).

[69] J.C. Ries, R.J. Eanes and M.M. Watkins, Confirming the Frame-Dragging Effect with Satellite Laser Ranging, *16th International Workshop on Laser Ranging*, 13-17 October 2008, Poznan, Poland. See also J.C. Ries, Relativity in Satellite Laser Ranging, American Astronomical Society, IAU Symposium 261. Relativity in Fundamental Astronomy: Dynamics, Reference Frames, and Data Analysis (Virginia Beach, VA, USA, 27 April - 1 May 2009).

[70] R. Koenig, B. Moreno Monge and G. Michalak, Some aspects and perspectives of measuring Lense-Thirring with GNSS and geodetic satellites. *Second International LARES Science Workshop*, Accademia dei Lincei, Rome, September 2012.

[71] I. Ciufolini, A. Paolozzi, E.C. Pavlis, J. Ries, R. Koenig, R. Matzner and G. Sindoni, The LARES Space Experiment: LARES Orbit, Error Analysis and Satellite Structure, in *John Archibald Wheleer and General Relativity* I. Ciufolini and R. Matzner eds., 371-434 (Springer Verlag, 2010).

[72] I. Ciufolini, A. Paolozzi, E.C. Pavlis, J. Ries, V. Gurzadyan, R. Koenig, R. Matzner, R. Penrose and G. Sindoni, Testing General Relativity and gravitational physics using the LARES satellite *The European Physical Journal Plus* **127**, 127 (2012).

[73] A. Paolozzi and I. Ciufolini, LARES successfully launched in orbit: Satellite and mission description, *Acta Astronautica* (2013).

[74] I. Ciufolini, B. Moreno Monge, A. Paolozzi, R. Koenig, G. Sindoni and G. Michalak, Monte Carlo Simulations of the LARES space experiment to test General Relativity and fundamental physics. To be published (2013). See also [75].

[75] B. Moreno Monge, R. Koenig, G. Michalak, I. Ciufolini, A. Paolozzi and G. Sindoni, Preliminary study for the measurement of the Lense- Thirring effect with the GALILEO satellites. To appear in *Acta Futura* (2013).

[76] M.R. Pearlman, J.J. Degnan and J.M. Bosworth, The International Laser Ranging Service *Advances in Space Research* **30**, 135-143, DOI:10.1016/S0273-1177(02)00277-6 (2002).

[77] S.W. Hawking and G.F.R. Ellis, *The Large Scale Structure of Space-Time* (Cambridge University Press, 1975).

[78] J.B. Hartle, *Gravity: An Introduction to Einsteins General Relativity* (Addison Wesley, San Francisco, 2003).

[79] W. Rindler, *Relativity: Special, General, and Cosmological* (Oxford University Press, Oxford, 2001).

[80] J. Ehlers, Survey of General Relativity Theory, in *Relativity, Astrophysics and Cosmology*, edited by W. Israel (Reidel Publishing) pp. 1-125 (1973).

[81] J. Ehlers and R. Geroch, Equation of motion of small bodies in relativity, *Ann. Phys.* **309**, 232 (2004).

[82] R. Geroch and P.S. Jang, Motion of a body in general relativit,y *J. Math. Phys.* **16**, 65 (1975).

[83] S. Zhu, Ch. Reigber and R. Koenig, Integrated Adjustment of CHAMP, GRACE, and GPS Data, *Journal of Geodesy* **78** 103-108 (2004).

[84] D.E. Pavlis et al., GEODYN operations manuals (Contractor Report, Raytheon, ITSS, Landover MD, 1998).

[85] C.F. Martin and D.P. Rubincam, Effects of Earth albedo on the LAGEOS I satellite, *J. Geophys. Res. B* **101**, 3215 (1996).

[86] D.P. Rubincam, Yarkovsky Thermal Drag on LAGEOS. *J. Geophys. Res. B* **93**, 13805 (1988).

[87] D.P. Rubincam, Drag on the LAGEOS satellite. *J. Geophys. Res.* **95** (B11), 4881-4886 (1990).

[88] I. Ciufolini, E.C. Pavlis and R. Peron, Determination of frame-dragging using Earth gravity models from CHAMP and GRACE, *New Astronomy* **11** 527-550 (2006).

arXiv:1504.07998v2 [math.DG] 18 May 2016

# ON THE QUADRATIC INVARIANT OF BINARY SEXTICS

MACIEJ DUNAJSKI AND ROGER PENROSE

Abstract. We provide a geometric characterisation of binary sextics with vanishing quadratic invariant.

## 1. Introduction

Classical invariant theory was formulated and developed by Cayley, Salmon, Sylvester, Young and others in the second half of the $19^{th}$ century. After a quiet period which lasted for the most part of the $20^{th}$ century, the theory has reappeared in algebraic geometry as the *geometric invariant theory* [13], in representation theory of $GL(2)$, and in other areas of modern mathematics [2]. Each of these areas replaces the classical terminology by its own language. In this paper we shall nevertheless follow the old–fashioned terminology of [8]. This is in line with other modern expositions of the subject [14, 12].

Despite the recent developments, some of the problems left over from the $19^{th}$ century remain open. One class of such problems has to do with finding the interpretation of the vanishing of invariants and covariants. This should be expressed in terms of the underlying projective geometry of roots on the Riemann sphere. The aim of this paper is to solve a problem from this category, and interpret the vanishing of the *apolar* invariant for a binary sextic, and more generally for binary quantics of even degree.

Consider a polynomial of degree six in $x$ with complex coefficients $\boldsymbol{\psi} = (\psi_0, \dots, \psi_6)$

$$\psi(x) = \psi_0 x^6 + 6\psi_1 x^5 + 15\psi_2 x^4 + 20\psi_3 x^3 + 15\psi_4 x^2 + 6\psi_5 x + \psi_6. \tag{1.1}$$

Substituting

$$x = \frac{a\tilde{x} + b}{c\tilde{x} + d}, \quad \text{where} \quad ad - bc = 1,$$

and multiplying the resulting expression by $(c\tilde{x} + d)^6$ to clear the denominators gives the polynomial $\tilde{\psi}(\tilde{x})$ with the coefficients $\tilde{\boldsymbol{\psi}}$ given by a linear transformation of $\boldsymbol{\psi}$. The function

$$\mathcal{I}(\boldsymbol{\psi}) = 2\psi_0\psi_6 - 12\psi_1\psi_5 + 30\psi_2\psi_4 - 20\psi_3^2 \tag{1.2}$$

is an *invariant* of the sextic, as $\mathcal{I}(\boldsymbol{\psi}) = \mathcal{I}(\tilde{\boldsymbol{\psi}})$.

Vanishing of any invariant of a binary quantic – the precise definitions are given in the next Section – describes some geometric property of the configurations of the roots of the quantic regarded as points on the two–dimensional sphere $\mathbb{CP}^1 = \mathbb{C} + \{\infty\}$. The analog of the quadratic invariant (1.2) can be constructed for any polynomial of even degree - see formula (2.4). In this paper we find a geometric interpretation of the condition $\mathcal{I} = 0$. This natural question does not seem to have been answered by the classical invariant theorists in the $19^{th}$ and early $20^{th}$ centuries, except when the the quantic has degree two or four: A generic quartic has four distinct roots, and the condition $\mathcal{I} = 0$ implies that their cross ratio is a cube root of unity. This is the equianharmonic condition. The roots of the quartic, when viewed as points on the Riemann sphere, can in this case be transformed into vertices of a regular tetrahedron by a Möbius transformation. The answer for binary sextics appears not to be known, and trying to understand this case in the context of twistor theory of $G_2$

*Date*: April 25, 2016.

structures [4, 3] is one motivation for this paper. It turns out that the sextic case can be reduced to the quartic in a sense made precise by the following Theorem

**Theorem 1.1.** *Let $X_1, X_2, X_3, X_4$ be four points on a two–dimensional sphere such that the stereographic projection of one of the roots of the sextic (1.1) from any of these four points lies in the centroid of the projections of the remaining five roots. Then $\mathcal{I}(\boldsymbol{\psi}) = 0$ if and only if the points $X_1, \dots, X_4$ can be transformed into vertices of a regular tetrahedron by a Möbius transformation (if they are distinct), or if at least three of these points coincide.*

First we will need to establish that there are indeed only four points, up to multiplicity, with the property stated in this Theorem. This, together with the rest of the proof with be presented in Section 3.

The invariant (1.2) is quadratic in the coefficients of the sextic, and thus given any five points on the sphere there will generically exist two points which complement these five points to roots of a sextic with $\mathcal{I} = 0$. In Proposition 4.2 we shall characterise the non–generic configurations of five points such that any choice of a distinct sixth point yields a sextic with $\mathcal{I} \neq 0$. It will be shown that any such non-generic configuration is projectively equivalent to vertices of a square pyramid.

The sextic case is special in some ways, but the general method in the paper together with an inductive argument applies to binary quantics of any even degree.

*Acknowledgements.* We thank Robert Bryant, Mike Eastwood, Nigel Hitchin and others for useful discussions.

## 2. Quantics and invariants

A binary quantic is a homogeneous polynomial in two variables which we shall call $(x, y)$. We shall consider binary quantics of even degree

$$\psi(x, y) = \sum_{k=0}^{2n} \binom{2n}{k} \psi_k x^{2n-k} y^k. \tag{2.3}$$

The coefficients of the quantic $\boldsymbol{\psi} = (\psi_0, \dots, \psi_{2n})$ are assumed to be complex numbers. There exists a unique, up to an overall scale, quadratic invariant

$$\mathcal{I}(\psi) = 2 \sum_{k=0}^{2n} (-1)^{2n-k} \binom{2n}{k} \psi_k \psi_{2n-k} \tag{2.4}$$

one can associate to the quantic[1]. The invariance of $\mathcal{I}$ is to be understood in the following way: Consider the linear action of $GL(2, \mathbb{C})$ on $\mathbb{C}^2$ given by the change of variables

$$x = a\tilde{x} + b\tilde{y}, \qquad y = c\tilde{x} + d\tilde{y}, \qquad ad - bc \neq 0.$$

Given a binary quantic $\psi(x, y)$, let $\widetilde{\psi}(\tilde{x}, \tilde{y})$ be a binary quantic given by

$$\begin{aligned} \widetilde{\psi}(\tilde{x}, \tilde{y}) &= \sum_{k=0}^{2n} \binom{2n}{k} \psi_k \, (a\tilde{x} + b\tilde{y})^{2n-k} \, (c\tilde{x} + d\tilde{y})^k \\ &= \tilde{\psi}_0 \tilde{x}^{2n} + 2n\tilde{\psi}_1 \tilde{x}^{2n-1}\tilde{y} + n(2n-1)\tilde{\psi}_2 \tilde{x}^{2n-2}\tilde{y}^2 + \dots + \tilde{\psi}_{2n}\tilde{y}^{2n}. \end{aligned}$$

This induces an irreducible embedding $GL(2, \mathbb{C}) \subset GL(2n+1, \mathbb{C})$, as the $(2n+1)$ coefficients of $\tilde{\psi}$ are linear homogeneous functions of the coefficients of $\psi$.

[1] Sylvester [18] calls it the *quadrinvariant*. In his latter works, see e.g. [19], he proposed an analogy between classical invariant theory and molecular chemistry. A binary sextic would correspond to an atom with six free valent electrons, and the quadratic invariant is then the bi-atomic molecule.

**Definition 2.1.** *A covariant of a binary quantic is a polynomial* $I = I(\psi_0, \ldots, \psi_{2n}, x, y)$ *such that*

$$I(\tilde{\psi}_0, \ldots, \tilde{\psi}_{2n}, \tilde{x}, \tilde{y}) = (ad - bc)^w I(\psi_0, \ldots, \psi_{2n}, x, y).$$

*The number* $w$ *is called the weight of the covariant. A covariant which only depends on the coefficients of the quantic, and not on* $(x, y)$ *is called an invariant.*

Thus the degree–two invariant (2.4) has weight $2n$. There are other invariants of degree higher than two [8]. In the case of the binary sextic there are four more invariants, of degree 4, 6, 10 and 15 respectively connected with $\mathcal{I}$ by a syzygy of degree 30.

2.1. **Transvectants.** Let $V_m = \mathrm{Sym}^m(\mathbb{C}^{2\vee})$ be the $(m+1)$–dimensional complex vector space of binary quantics of degree $m$, where $\mathbb{C}^{2\vee}$ is the dual of $\mathbb{C}^2$. Given two binary quantics $\phi \in V_n$ and $\psi \in V_m$, the $k$th transvectant is a map $< \, , \, >_k \colon V_m \times V_n \to V_{m+n-2k}$ given by a quantic of degree $n + m - 2k$

$$< \phi, \psi >_k := \sum_{j=1}^{k} (-1)^j \binom{k}{j} \frac{\partial^k \phi}{\partial x^{k-j} \partial y^j} \frac{\partial^k \psi}{\partial x^j \partial y^{k-j}}. \tag{2.5}$$

Thus, for any $k \leq \min(m, n)$, transvectants are covariants of weight $k$ and degree two. One of the results in the classical invariant theory is that all covariants and invariants arise from the transvectants operations [8, 14].

**Definition 2.2.** *The quantic* $\phi \in V_n$ *is apolar to* $\psi \in V_m$ *where* $m \geq n$ *if* $< \psi, \phi >_n = 0$.

In the special case $n = m$ the apolarity condition is given by

$$\sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} \psi_k \phi_{n-k} = 0.$$

Any quantic of an odd degree is apolar to itself. A quantic of an even degree is apolar to itself iff the quadratic invariant $\mathcal{I}$ given by (2.4) vanishes.

## 3. Characterisation of $\mathcal{I}$

Let $\psi$ be a sextic (1.1) which is generic in the sense that its six complex roots are distinct, and let $\mathcal{I}(\psi)$ be given by (1.2).

**Proposition 3.1.** *Let* $P_1, \ldots, P_6$ *be six points on the sphere* $\mathbb{CP}^1$ *corresponding to the roots of a sextic* $\psi$. *Then* $\mathcal{I}(\psi) = 0$ *if and only if the four roots of the quartic* $< \kappa, \rho >_1$ *can be transformed into vertices of a regular tetrahedron (if they are distinct) or contain a root of multiplicity at least three. Here* $\kappa$ *and* $\rho$ *are any quintic and a linear form respectively such that* $\psi(x) = \kappa(x)\rho(x)$.

**Proof.** Any quintic corresponding to distinct points $P_1, \ldots, P_5$ associates four points on the sphere to a given point $P_6$, such that the quartic defining the four points is a transvectant of the quintic with a linear form corresponding to $P_6$. Given $\psi(x, y) = \kappa(x, y)\rho(x, y)$, and using $< \rho, \rho >_1 = 0$ we compute

$$\mathcal{I}(\psi) = < \psi, \psi >_6 = - < \delta, \delta >_4 = -\mathcal{I}(\delta), \quad \text{where} \quad \delta = < \kappa, \rho >_1 .$$

Thus $\mathcal{I}(\psi) = 0$ iff $\mathcal{I}(\delta) = 0$, where $\delta$ is the quartic defined above. But $\mathcal{I}(\delta) = 0$ if at least three of the roots of $\delta$ coincide, or if the cross ratio of the four distinct roots is a cube root of unity (the equianharmonic case). For completeness we give the proof of this fact, well known in the $19^{th}$ century literature. Set $y = 1$, consider a general quartic

$$\delta = \delta_0 x^4 + 4\delta_1 x^3 + 6\delta_2 x^2 + 4\delta_3 x + \delta_4$$

with four distinct roots. This has $\mathcal{I}(\delta) = 2\delta_0\delta_4 - 8\delta_1\delta_3 + 6(\delta_2)^2$. Now consider the quartic corresponding to a regular tetrahedron with one vertex at $\infty$. This quartic is represented by $(x-1)(x-\omega)(x-\omega^2)$ where $\omega^3 = 1$, and we find $\mathcal{I} = 0$. The cross ratio of the roots is given by $\omega$, and conversely any quartic with the cross ratio of four roots given by a cube root of unity is projectively equivalent to the tetrahedral quartic.

□

We can now give the proof of the result stated in the Introduction. In the proof, and in the remaining part of the paper we shall write

$$\psi = P_1 P_2 \dots P_m \tag{3.6}$$

to denote a binary quantic defined (up to an overall non–zero multiple) by its roots corresponding to points $P_1, P_2, \dots, P_m$ on the sphere.

**Proof of Theorem 1.1.** Consider the quartic $\delta =< \kappa, \rho >_1$ introduced in Proposition 3.1. Let its four roots correspond to the (not necessarily distinct) points $X_1, \dots, X_4$. Thus, using the notation (3.6), the quartic $\delta$ is apolar (in the sense of Definition 2.2) to a quartic $XXXX$, where $X$ is any of the roots $X_i$.

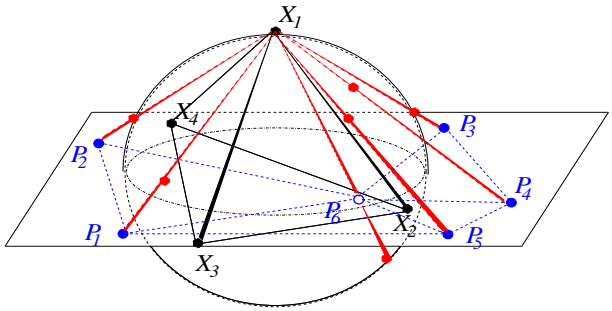



**Figure 1.** *Stereographic projection of $P_6$ to the centroid of $P_1, \dots, P_5$.*

Equivalently the quintic $\kappa = P_1P_2P_3P_4P_5$ is apolar to the quintic $\chi = P_6X_iX_iX_iX_i$ for any $i = 1, \dots, 4$. This happens if and only if the stereographic projection of $P_6$ from any of the four points $X_i$ lies in the centroid of the projections of the five points $P_1, \dots, P_5$. This fact was observed in [16, 20]. Indeed, if the homogeneous coordinates of the projected points $P_i$ are $(1, -x_i)$ where $i = 1, \dots, 6$ and the coordinates of the north pole $X$ are $(0, 1)$ then $< P_i, X >= 1, < P_i, P_j >= x_i - x_j$, so

$$< \kappa, \chi >_5 = \sum_{i=1}^{5} (x_i - x_6) = 0, \qquad \text{and} \quad x_6 = \frac{1}{5}\sum_{i=1}^{5} x_i.$$

We note that the centroid of the set of points is not invariant under the projective transformations, but is invariant under the subgroup of affine transformations $x \to \alpha x + \beta$ preserving the north pole of the stereographic projection. The statement in the Theorem 1.1 is nevertheless projectively invariant.

□

This proof extends to binary quantics of arbitrary even degree. Assume we know a geometrical interpretation for vanishing of $\mathcal{I}$ for a binary quantic $\psi_{2n-2}$ of degree $2n-2$. We can now characterise the self–apolarity of a binary quantic $\psi_{2n}$ of degree $2n$ by the following inductive argument: Given $2n-1$ distinct points $(P_1, \dots, P_{2n-1})$ on the sphere, let $\kappa$ be the unique (up to a non–zero multiple) binary quantic of degree $(2n-1)$ with these points corresponding to its roots. The points $(P_1, \dots, P_{2n-1}, P_{2n})$ are roots of a binary quantic $\psi_{2n}$ with $\mathcal{I}(\psi_{2n}) = 0$ if a linear form $\rho$ with a root corresponding to the point $P_{2n}$ is such that the binary quantic $\psi_{2n-2} =< \kappa, \rho >$ has $\mathcal{I}(\psi_{2n-2}) = 0$.

Given $2n$ unordered points $\psi_{2n} = \{P_1, P_2, \ldots, P_{2n}\}$ on the sphere, split them into a set of $2n-1$ points $\kappa = \{P_1, P_2, \ldots, P_{2n-1}\}$ together with one point $\{P_{2n}\}$. Let $\delta = \{X_1, X_2, \ldots, X_{2n-2}\}$ be the points on the sphere such that the stereographic projection of $P_{2n}$ from any $X_i$ is the centroid of the stereographic projections of $P_1, \ldots, P_{2n-1}$. Then the set of points $\psi$ is self-apolar iff the set of points $\delta$ is self–apolar. The set $\delta$ consist of the roots of the polynomial equation

$$< \kappa, \chi >_{2n-1} = 0, \quad \text{where,} \quad \chi = P_{2n} X X \ldots X \tag{3.7}$$

counted with multiplicity.

*Example.* The multiplicities of the elements of $\delta$ can depend on the choice of the point $P_{2n}$ from the set $\psi$. Consider the sextic

$$\psi = (x-1)(x-\omega)(x-\omega^2)(x-\omega^3)(x-\omega^4), \quad \text{where} \quad \omega^5 = 1 \tag{3.8}$$

corresponding to a pentagonal pyramid with one root placed at $\infty$. This has $\mathcal{I} = 0$. Thus $\psi = \{1, \omega, \omega^2, \omega^3, \omega^4, \infty\}$. Taking $P_6 = \infty$ gives the quartic (3.7) to be $x^4$ which has one quadrupole root $x = 0$ and thus is self–apolar. Choosing instead $P_6 = 1$ gives the polynomial (3.7)

$$x^4 + 6x^3 + 6x^2 + 6x + 6 = 0$$

which has four distinct roots with equianharmonic cross ratio.

3.1. **Canonical form.** Almost all sextics can be put in the Sylvester Canonical Form [18]

$$\psi = Cu^6 + Av^6 + Bw^6 + uvw(u-v)(v-w)(w-u), \tag{3.9}$$

where $u + w + v = 0$ - thus to obtain (1.1) set $u = x, v = 1, w = -x-1$ in this formula[2]. In this form the quadratic invariant (1.2) is

$$\mathcal{I}(\boldsymbol{\psi}) = 2CA + 2CB + 2BA - 2.$$

If the sextic does not have a root at $\infty$ we can assume that $C + B \neq 0$, and solve for $A = (1 - CB)/(C + B)$. This gives a canonical form of a generic self-apolar sextic. It parametrises a non–singular open orbit in the space of all self-apolar sextics. Setting $B = (b+c)/(3b-3c), C = (6-b-c)/(3b-3c)$ gives

$$\psi = x^6 + 2bx^5 + 5bx^4 + \frac{20}{6}(b+c)x^3 + 5cx^2 + 2cx + \frac{1}{36}(b+c)^2 + \frac{1}{4}(b-c)^2, \tag{3.10}$$

so the self-apolar sextic depends on two parameters. In general the binary quantic of degree $2n$ with $\mathcal{I} = 0$ depends on $2n-4$ arbitrary parameters, up to the Möbius transformation.

3.2. **Catalectant.** We shall finish off this section giving an alternative interpretation of the condition $\mathcal{I} = 0$, which brings up another quartic invariant of binary sextics.

The seven–dimensional complex vector space of binary sextics $V_6$ belongs to the space of endomorphisms of the four–dimensional space of binary cubics $V_3$, where the endomorphism corresponding to a sextic $\psi$ is given by the transvectant $\phi \rightarrow < \psi, \phi >_3$, where $\phi \in V_3$. Consider a complex eigenvalue problem

$$< \psi, \phi >_3 = \lambda \phi. \tag{3.11}$$

The eigenvalue $\lambda$ has weight three under the $GL(2)$ action on $V_6$. It is also an eigenvalue, in the ordinary sense, of a 4 by 4 matrix corresponding to $\psi$ by choosing a basis of $V_3$. If

[2]The exceptional sextics which can not be put in the canonical form are classified in [6].

$\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are the four eigenvalues of $\psi$ then the characteristic polynomial is a quartic[3]

$$\begin{aligned} \chi_\psi(\lambda) &= 8(\lambda-\lambda_1)(\lambda-\lambda_2)(\lambda-\lambda_3)(\lambda-\lambda_4) \\ &= 8\lambda^4 + 4\mathcal{I}\,\lambda^2 - \mathcal{J}. \end{aligned} \tag{3.12}$$

Such quartic is canonically associated to every sextic, and has weight 12. The degree four invariant $\mathcal{J} =<< \psi, \psi >_3, < \psi, \psi >_3>_6$ appearing in (3.12) is the *catalectant* of the sextic. Its zero set is the closure of the locus of sextics expressible as the sum of three 6th powers [7]. Equivalently, the catalectant vanishes if the sextic admits an apolar cubic [8]. The latter result follows directly from setting $\lambda = 0$ in (3.12). If $\mathcal{I} = 0$ and $\mathcal{J} \neq 0$, then the roots of (3.12) form a *harmonic* set. This corresponds to a square on an equator in $\mathbb{CP}^1$.

## 4. Maximally separated quintics

In this section we shall consider a problem of recovering a sextic $P_1P_2\dots P_6$ with $\mathcal{I} = 0$ from a quintic. Let us project stereographically the sphere to the complex plane from one of the roots - say $P_6 = \infty$. Given four points $P_1, \dots, P_4$ on the plane, we can now look for a fifth point $P_5$ such that $\mathcal{I} = 0$. Rewriting the sextic (1.1) with the root corresponding to $P_6$ at $\infty$ as

$$\psi = (x-x_1)(x-x_2)(x-x_3)(x-x_4)(x-x_5)$$

and comparing the coefficients of various powers of $x$ we find that

$$\mathcal{I} = a\, {x_5}^2 + 2b\, x_5 + c, \tag{4.13}$$

where $(a, b, c)$ are polynomials in $(x_1, \dots, x_4)$. This expression has at most two roots $x_5 \neq \infty$, thus given five distinct points on $\mathbb{CP}^1$ there exist at most two points such that the invariant $\mathcal{I}$ of the sextic defining the resulting six distinct points vanishes. Generically, if $a \neq 0$, there will be two such points.

*Example.* Consider a quintic corresponding to four points on the base of an equatorial regular pentagon, and a point at infinity. Equation (4.13) gives two possibilities for the sixth s.t. $\mathcal{I} = 0$, one of which is gives rise to a pentagonal pyramid (3.8)(Figure 2).

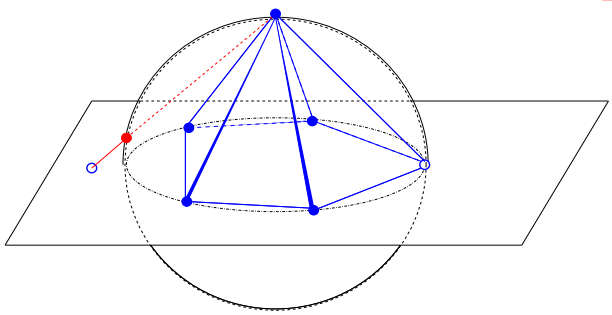


**Figure 2**. *Regular pentagon with $x_5 = 1$ and its companion with $x_5 = -3$.*

The next example is of a non-generic type

*Example–Regular Octahedron.* Assume that the six points $P_i$ form the vertices of the regular octahedron, and project from one of its vertices $P_6$ which does not belong to the square. The resulting sextic is $\psi = x^5 - x$, and $\mathcal{I} = 1/3$. We can transform the four corners of the square $P_1, \dots, P_4$ to $\pm 1, \pm i$, and look for a point $P_5$ such that $\mathcal{I} = 0$. The resulting sextic is

$$(x^4 - 1)(x - x_5)$$

[3]This is the equation **R** on page 24 in [18], used to construct the canonical form (3.9).

and we find that $\mathcal{I} = 1/3$ for any value of $x_5$. Thus there is no $P_5$ which complements the five vertices of the square pyramid to an octahedron with $\mathcal{I} = 0$. This example corresponds to $a = b = 0, c \neq 0$ in (4.13).

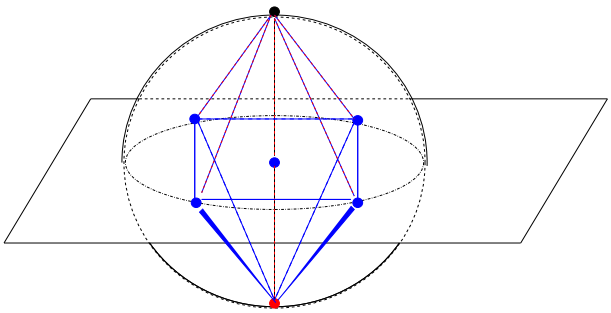


**Figure 3.** *Five maximally separated points.*

**Definition 4.1.** *We shall call five distinct points on the sphere maximally separated if $\mathcal{I} \neq 0$ for the sextic with six distinct roots resulting from any choice of the sixth point.*

**Proposition 4.2.** *Any maximally separated five points on the sphere can be transformed into vertices of a square pyramid by a Möbius transformation.*

**Proof.** Stereographicaly project from any of the five maximally separated points. The remaining four points are the distinct roots of a quartic

$$\gamma(x) = x^4 + 4b_1x^3 + 6b_2x^2 + 4b_3x + b_4.$$

Now consider the sextic $\psi$ with one root at infinity given by $\psi = \gamma(x)(x - x_5)$. Computing (4.13) and equating $a$ and $b$ to zero gives $b_2 = b_1^2, b_3 = b_1^3$, and now $\mathcal{I} \neq 0$ unless all four roots of $\gamma$ coincide which we have excluded. The roots of the resulting quartic are of the form $\alpha \pm \beta, \alpha \pm i\beta$, where $\alpha$ and $\beta$ are complex numbers depending on $b_1$ and $b_4$. These roots are *harmonically separated*, with cross ratio equal to $-1, 2$ or $1/2$. Thus the corresponding points can be transformed to vertices of a square. We can use the remaining freedom in the Möbius transformations to set two roots to $\pm i$. Then either $b_1 = 0$ in which case the remaining two roots are at $\pm 1$, or $b_1 = \pm 1$ in which case the roots are at $2 \pm i$ or $-2 \pm i$. Both squares are equivalent under the Mobiüs transformations. Thus the resulting five points form a square pyramid (Figure 3).

□

## 5. Motivation: Twistor theory of $G_2$ structures

We shall close the paper explaining the motivation of characterising binary sextics with $\mathcal{I} = 0$ coming from twistor theory of $G_2$ structures. For the sake of the following discussion twistor theory is a correspondence [15] between global algebraic geometry of curves in complex two–folds or three-folds, and local differential geometry on the moduli spaces of these curves.

Let $\mathcal{Z}$ be a complex two–fold with a family of rational curves $L_m \cong \mathbb{CP}^1$ parametrised by points $m \in M$, where $M$ is some complex manifold. For any $m$, the embedding of $L_m$ in $\mathcal{Z}$ is, to the first order, described by the normal bundle $N(L_m) = T\mathcal{Z}/TL_m$. This is a holomorphic line bundle $\mathcal{O}(k)$, for some integer $k$ which we shall assume to be positive and even. The obstruction group $H^1(L_m, N(L_m)) = 0$ vanishes, and the Kodaira deformation theorem [11] states that there exists a canonical isomorphism

$$T_mM \cong H^0(L_m, N(L_m)) = \mathrm{Sym}^k(\mathbb{C}^{2\vee})$$

between vectors tangent to $M$ and binary quantics of degree $k$. This is where the quadratic invariant (2.4) becomes relevant: if $k$ is even, we can define a holomorphic conformal structure $[g]$ on $M$ by declaring a vector field $U \in \Gamma(TM)$ to be null iff the corresponding quantic has $\mathcal{I}_2(U) = 0$. This is a quadratic condition, and the holomorphic light-cone of any point in $M$ is a surface of co–dimension one in $M$, so $[g]$ is indeed well defined.

Let us now restrict to the case $k = 6$, where vector fields correspond to binary sextics, and $\dim_{\mathbb{C}} M = 7$, [4] . In this case there exists a skew–symmetric three–form $\Psi \in \Lambda^3(M)$ given by

$$\Psi(U, V, W) = << U, V >_3, W >_3$$

where $<,>_3$ is the third transvectant (2.5) of two binary sextics, and we use the same symbols to denote vector fields and corresponding sextics. This three–form is compatible with $[g]$ is a sense that

$$(U \lrcorner \Psi) \wedge (U \lrcorner \Psi) \wedge \Psi = 0, \quad \text{iff} \quad \mathcal{I}_2(U) = 0.$$

The invariants $\mathcal{I}_2$ and $\Psi$ have weights six and nine respectively, so changing a metric $g \in [g]$ yields $g \to \Omega^6 g, \Psi \to \Omega^9 \Psi$ where $\Omega$ is a non–vanishing function on $M$. Therefore the structure group of $TM$ reduces to $GL(2, \mathbb{C}) \subset {G_2}^{\mathbb{C}} \times \mathbb{C}$, where ${G_2}^{\mathbb{C}}$ is the complexification of the exceptional Lie group $G_2$.

There are only few known examples of this construction which lead to positive–definite $G_2$ structures on Riemannian manifolds $M_{\mathbb{R}}$. Bryant's weak $G_2$ holonomy metric [1] on $M_{\mathbb{R}} = SO(5)/SO(3)$ arises from a family of $Sp(4)$ invariant rational sextics [5]. Another example corresponds to $\mathcal{Z} = \mathbb{CP}^2$, and $M$ being the homogeneous space $PSL(3, \mathbb{C})/\mathbb{C}^*$ of ternary cuspidal cubics in $\mathcal{Z}$. The cuspidal cubics are rational, but singular. The Kodaira theory nevertheless applies as the contact lifts of the cuspidal cubics to $T\mathbb{CP}^2$ are smooth. There exist three real slices of $M$, one of which is $M_{\mathbb{R}} = SU(2, 1)/U(1)$. In [3] it is shown that $M_{\mathbb{R}}$ admits a $G_2$ structure which is co–calibrated [3].

## 6. Conclusions

We have found a geometric interpretation of vanishing of the quadratic invariant associated with a binary sextic, and more generally with any binary quantic of even degree. The result should have an interpretation in the theory of hyperelliptic curves, as projective equivalence classes of binary sextics with distinct roots correspond to points on the moduli space of genus two algebraic curves. Igusa [10] considered a zero locus of various invariants for sextics, but the case of vanishing quadratic invariant has not been analysed in this work.

The problem addressed in this paper can also be reformulated in the context of the representation theory of $SO(3, \mathbb{C})$ which is related to $SL(2, \mathbb{C})$ by the homomorphism $SL(2, \mathbb{C})/\mathbb{Z}_2 \cong SO(3, \mathbb{C})$. The isomorphism $\mathbb{C}^3 = \mathrm{Sym}^2(\mathbb{C}^{2\vee})$ identifies complex vectors in $\mathbb{C}^3$ with symmetric 2 by 2 matrices with complex coefficients. The *null* vectors correspond to rank one matrices, and this gives rise to a holomorphic conformal structure on $\mathbb{C}^3$. The seven–dimensional space $\mathrm{Sym}^6(\mathbb{C}^{2\vee})$ of binary sextics is identified with a subspace of $\mathrm{Sym}^3(\mathbb{C}^{3\vee})$ which consist of harmonic ternary cubics, i. e. those forms $\Psi_{ijk} Z^i Z^j Z^k, i, j, k = 1, \ldots, 3$ which satisfy $\delta^{ij} \Psi_{ijk} = 0$. Hitchin [9] showed that a generic harmonic ternary cubic in $\mathbb{CP}^2$ passes through two sets of six points corresponding to six axes of a regular icosahedron in $\mathbb{C}^3$. In this formulation the quadratic invariant is given by the norm of the cubic $\Psi_{ijk}$ taken w. r. t. the conformal structure defined above. It is however not clear what is the geometric meaning of its vanishing in terms of Hitchin's icosahedron.

## Appendix: Two component spinors

A convenient way to represent binary quantics and the associated invariants uses the two–component spinor notation [17]. Let the capital letters $A, B, \ldots$ denote indices taking values 0 and 1. The general quantic (2.3) is represented by a *symmetric spinor* of valence $2n$. The Fundamental Theorem of Algebra states that any such spinor factorises into valence one spinors

$$\psi_{AB\cdots C} = \alpha_{(A}\beta_B \ldots \gamma_{C)},$$

where the round brackets on the RHS denote the symmetrisation. The binary quantic (2.3) is then given by

$$\begin{aligned} \psi &= \psi_{AB\ldots C}\pi^A\pi^B\ldots\pi^C \\ &= (\alpha_0 x + \alpha_1 y)(\beta_0 x + \beta_1 y)\ldots(\gamma_0 x + \gamma_1 y) \sim (x - x_1)(x - x_2)\ldots(x - x_{2n}), \end{aligned}$$

where $\pi^0 = x, \pi^1 = y$ and $\psi_0 = \psi_{00\ldots0}, \quad \psi_1 = \psi_{10\ldots0}, \quad \ldots, \quad \psi_{2n} = \psi_{11\ldots1}$. Thus the complex numbers $x_1 = -\alpha_1/\alpha_0, x_2 = -\beta_1/\beta_0, \ldots, x_{2n} = -\gamma_1/\gamma_0$ are the roots of the inhomogeneous polynomial of degree $2n$ obtained by setting $y = 1$. The invariant (2.4) is in this notation given by

$$\mathcal{I} = \psi_{AB\ldots C}\psi^{AB\ldots C},$$

where the indices are lowered by the anti–symmetric matrix $\varepsilon_{AB}$ with $\varepsilon_{01} = 1$, so that

$$\psi_{AB\ldots C} = \psi^{PQ\ldots R}\varepsilon_{PA}\varepsilon_{QB}\ldots\varepsilon_{RC}.$$

The $k$th transvectant (2.5) is

$$< \psi, \phi >_k = \varepsilon^{A_1B_1}\ldots\varepsilon^{A_kB_k}\psi_{A_1\ldots A_kA_{k+1}\ldots A_m}\phi_{B_1\ldots B_kB_{k+1}\ldots B_n}\pi^{A_{k+1}}\ldots\pi^{A_m}\pi^{B_{k+1}}\ldots\pi^{B_n}.$$

Using the spinor notation gives a simple proof of the following algebraic interpretation of the condition $\mathcal{I} = 0$

**Lemma .1.** *An even degree quantic $\psi \in V_{2n}$ with distinct roots is a sum of $(2n)^{th}$ powers of its factors iff $\mathcal{I}(\psi) = 0$.*

**Proof.** Let $\alpha_A, \beta_A, \ldots, \gamma_A$ be homogeneous coordinates of the points in $\mathbb{CP}^1$ corresponding to the roots of $\psi$ so that the condition stated in the Lemma becomes[4]

$$\alpha_{(A}\beta_B\ldots\gamma_{C)} = c_1\,\alpha_A\alpha_B\ldots\alpha_C + c_2\,\beta_A\beta_B\ldots\beta_C + \cdots + c_{2n}\,\gamma_A\gamma_B\ldots\gamma_C, \tag{A1}$$

where $c_1, c_2, \ldots, c_{2n}$ are some constants determined by the roots. The condition $\mathcal{I} = 0$ is then equivalent to

$$\frac{\partial^{2n}\psi}{\partial\alpha\partial\beta\ldots\partial\gamma} = 0, \quad \text{where} \quad \frac{\partial}{\partial\alpha} := \alpha^A\frac{\partial}{\partial\pi^A} \text{ etc.}$$

Thus

$$\frac{\partial^{2n-1}\psi}{\partial\beta\ldots\partial\gamma} = \tilde{c}_1(\alpha_A\pi^A)$$

for some constant $\tilde{c}_1$, as both sides are homogeneous of degree one in $\pi^A$. The successive $(2n-2)$ integrations yield (A1), with $c_1 = (2n)!^{-1} < \alpha, \beta >_1{}^{-1}\ldots < \alpha, \gamma >_1{}^{-1}(\tilde{c}_1)$ etc.

□

[4]Formula (A1) holds for any binary quantic of odd degree, as then $\mathcal{I}$ vanishes identically.

## References


[1] Bryant, R. L. (1987) Metrics with exceptional holonomy. Ann. of Math. **126**, 525–576.

[2] Dolgachev, I. (2003) *Lectures on Invariant Theory.* CUP.

[3] Doubrov, B. and Dunajski, M. (2012) Co–calibrated $G_2$ structure from cuspidal cubics. Ann. Global Anal. Geom. **42**, 247-265.

[4] Dunajski, M. and Godliński, M. (2012) $GL(2,\mathbb{R})$ structures, $G_2$ geometry and twistor theory. Quart. J. Math **63**, 101-132.

[5] Dunajski, M. and Sokolov, V. V. (2011) On 7th order ODE with submaximal symmetry. J. Geom. Phys. **61**, 1258-1262.

[6] Eastwood, M. G. and Isaev, A. V. (2013). Extracting Invariants of Isolated Hypersurface Singularities from their Moduli Algebras. Math. Ann. **356**, 73-98.

[7] Elliott E. B. (1895) *An introduction to the algebra of quantics.* OUP, Clarendon Press, Oxford.

[8] Grace, J. H. and Young, A. (1903). *The algebra of invariants.* CUP, Cambridge.

[9] Hitchin, N. (2010) Vector Bundles and the Icosahedron. Contemp. Math., **522**, Amer. Math. Soc., Providence, RI, 71-87.

[10] Igusa, J. I. (1960) Arithmetic variety of moduli of genus two. Ann. of Math. **72**. 612-649.

[11] Kodaira, K. (1963) On stability of compact submanifolds of complex manifolds, Am. J. Math. **85**, 79-94.

[12] Kung, J. P. S. and Rota, G. (1984) The invariant theory of binary forms. Bull. AMS **10**, 27-85.

[13] Mumford, D., Fogarty, J. and Kirwan, F. (1994) *Geometric invariant theory.* Springer-Verlag,

[14] Olver, P. (1999) *Classical invariant theory*, CUP, Cambridge.

[15] Penrose, R. (1976) Nonlinear gravitons and curved twistor theory, Gen. Rel. Grav. **7**, 31-52.

[16] Penrose, R. (1993) Orthogonality of General Spin States. Twistor Newsletter **36.**

[17] Penrose, R. and Rindler, W. (1987, 1988) *Spinors and space-time. Two-spinor calculus and relativistic fields.* Cambridge Monographs on Mathematical Physics. CUP, Cambridge.

[18] Sylvester, J. J. (1854) On the calculus of forms, otherwise the theory of invariants. Cambridge and Dublin Mathematical Journal IX, 85-103.

[19] Sylvester, J. J. (1878) On an Application of the New Atomic Theory to the Graphical Representation of the Invariants and Covariants of Binary Quantics, with Three Appendices. Am. Journ. Math. I. 64-125.

[20] Zimba, J. and Penrose, R. (1993) On Bell nonlocality without probabilities: more curious geometry. Stud. Hist. Philos. Sci. **24** , 697-720.



Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, UK.

*E-mail address*: `m.dunajski@damtp.cam.ac.uk`

The Mathematical Institute, University of Oxford,, Andrew Wiles Building,, Woodstock Road, Oxford OX2 6GG, UK.

*E-mail address*: `penroad@wadh.ox.ac.uk`

arXiv:1603.09674v1 [gr-qc] 29 Mar 2016

# A Test of General Relativity Using the LARES and LAGEOS Satellites and a GRACE Earth's Gravity Model

## Measurement of Earth's Dragging of Inertial Frames

Ignazio Ciufolini*[1,2], Antonio Paolozzi[2,3], Erricos C. Pavlis[4], Rolf Koenig[5], John Ries[6], Vahe Gurzadyan[7], Richard Matzner[8], Roger Penrose[9], Giampiero Sindoni[10], Claudio Paris[2,3], Harutyun Khachatryan[7] and Sergey Mirzoyan[7]

[1]Dip. Ingegneria dell'Innovazione, Università del Salento, Lecce, Italy
[2]Museo della fisica e Centro studi e ricerche Enrico Fermi, Rome, Italy
[3]Scuola di Ingegneria Aerospaziale, Sapienza Università di Roma, Italy
[4]Joint Center for Earth Systems Technology (JCET), University of Maryland, Baltimore County, USA
[5]Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences, Potsdam, Germany
[6]Center for Space Research, University of Texas at Austin, Austin, USA
[7]Center for Cosmology and Astrophysics, Alikhanian National Laboratory and Yerevan State University, Yerevan, Armenia
[8]Theory Center, University of Texas at Austin, Austin, USA
[9]Mathematical Institute, University of Oxford, Oxford, UK
[10]DIAEE, Sapienza Università di Roma, Rome, Italy

April 1, 2016

We present a test of General Relativity, the measurement of the Earth's dragging of inertial frames. Our result is obtained using about 3.5 years of laser-ranged observations of the LARES, LAGEOS and LAGEOS 2 laser-ranged satellites together with the Earth's gravity field model GGM05S produced by the space geodesy mission GRACE. We measure $\mu = (0.994 \pm 0.002) \pm 0.05$, where $\mu$ is the Earth's dragging of inertial frames normalized to its General Relativity value, 0.002 is the 1-sigma formal error and 0.05 is the estimated systematic error mainly due to the uncertainties in the Earth's gravity model GGM05S. Our result is in agreement with the prediction of General Relativity.

*ignazio.ciufolini@unisalento.it

# 1 Introduction

About one hundred years ago Albert Einstein completed the publication of a series of fundamental papers describing the gravitational theory known as General Relativity (GR) [1, 2, 3, 4]. Since then Einstein's gravitational theory has had experimental and theoretical triumphs, including the prediction and observation of the expansion of the universe, of black holes, gravitational lensing and gravitational waves [5, 6, 5, 7, 8, 9, 10, 11]. GR has today a number of practical applications to our everyday life [12] including its corrections that enable the Global Navigation Satellite System to reach accuracies at the level of a few decimeters [13].

Nevertheless, GR has not been reconciled with the other fundamental theory of modern physics: Quantum Mechanics. Further, Einstein's gravitational theory predicts the occurrence of spacetime singularities where every known physical theory ceases to be valid, the spacetime curvature diverges and time ends [14]. In 1998 observations of distant supernovae of type Ia implied the quite surprising result that the universe has an accelerated expansion [15, 16]. An explanation for this mysterious result can be found in the cosmological constant introduced by Einstein to avoid a dynamical universe and later, in 1931, abandoned by Einstein himself. However, the cosmological constant corresponds to vacuum energy and quantum field theory predicts that the vacuum energy should have a value approximately $10^{122}$ times larger than the *dark energy* [17, 18] density that is observed in the universe. To explain the accelerated expansion of the universe, dark energy should compose more than 70% of our universe, but its real nature is unknown. Other explanations include a time dependent vacuum energy with the exotic name of quintessence, and modifications of GR such as the so-called f(R) theories. Therefore, in spite of its experimental triumphs, Einstein's gravitational theory continues to need further accurate tests at all scales from solar system tests to astrophysical and cosmological observations.

Successful tests [8, 9, 10] of effects and phenomena predicted by GR include the well known perihelion precession of Mercury (and in general the periastron advance of an orbiting body), the equivalence principle and the time-dilation of clocks in a gravitational field, the deflection and time-delay of electromagnetic waves by a mass, the dynamics of the Moon, accurately measured by Lunar Laser Ranging and of binary pulsars [19, 20, 21], gravitational lensing and other relevant astrophysical observations. Gravitational waves have been indirectly observed at the level predicted by GR from the rate of change of the orbital period of the binary pulsar PSR B1913+16 [19]. Recently the two LIGO advanced detectors (Caltech and MIT) have directly detected the gravitational waves from the inspiral and merger of a pair of black holes [11] marking the beginning of the gravitational-wave astronomy.

# 2 Dragging of Inertial Frames

Among the intriguing phenomena predicted by GR, and so far only tested with approximately 10% accuracy, is the "dragging of inertial frames", or "frame-dragging" as

Einstein named it in 1913 [22]. Frame-dragging has relevant astrophysical applications to the dynamics of matter falling into rotating black holes and of jets in active galactic nuclei and quasars [23].

A test-gyroscope is a small current of mass in a loop and may be realized using a sufficiently small spinning top. In GR a gyroscope determines the axes of local nonrotating inertial frames. In such frames the equivalence principle holds so that the gravitational field is locally unobservable and all the laws of physics are the laws of Special Relativity theory. However in GR a gyroscope has a potential behaviour different from that in classical Galilei-Newton mechanics. In classical mechanics, a torque-free gyroscope is predicted to always point towards the same distant "fixed" stars. In contrast, in GR, a gyroscope is dragged by mass currents, such as the spinning Earth, and therefore its orientation can change with respect to the distant "fixed" stars. If we were to rotate with respect to the gyroscope, we would feel centrifugal forces, even though we may not rotate at all with respect to distant "fixed" stars [8].

Frame-dragging of a gyroscope is formally similar to the change of orientation of a magnetic dipole by a magnetic field generated by an electric current in electrodynamics [23]. In GR, a current of mass generates an additional contribution to the gravitational field, called gravitomagnetic field because of its formal analogy with electrodynamics. The gravitomagnetic field then exerts a torque on a gyroscope in the same way a magnetic field torques a magnetic needle in electrodynamics.

In 1918, Lense and Thirring [24] published the equations of the frame-dragging perturbations of the orbital elements of a satellite in the weak gravitational field of a slowly rotating body. The rate of change of the nodal longitude of the satellite, known as Lense-Thirring effect, is given by $\dot{\mathbf{\Omega}} = \frac{2\mathbf{J}}{a^3\,(1-e^2)^{3/2}}$, where $\mathbf{\Omega}$ is the nodal longitude of the satellite, $a$ its semimajor axis, $e$ its orbital eccentricity and $\mathbf{J}$ is the angular momentum of the rotating body. We recall that the node, ascending or descending, of a satellite is defined as the intersection of its orbit with the equatorial plane of the central body, in our case the Earth [25].

Frame-dragging was observed [26] in 1997-1998 by using the LAGEOS (LAser GEOdynamics Satellite) and LAGEOS 2 laser-ranged satellites [27] and measured with approximately 10% accuracy [28, 29, 30, 31] in 2004-2010, using LAGEOS, LAGEOS 2 and the Earth's gravity field determinations by the space geodesy mission GRACE [32, 33]. In 2011 the dedicated space mission Gravity Probe B, launched in 2004 by NASA, reported also a test of frame-dragging with approximately 20% accuracy [34].

LAGEOS was launched in 1976 by NASA, and LAGEOS 2 in 1992 by ASI and NASA [27]. They are two almost identical passive satellites covered with 426 corner cube reflectors to reflect back the laser pulses emitted by the stations of the Satellite Laser Ranging (SLR) network [35]. SLR allows measurement of the position of the LAGEOS satellite with an accuracy that can reach a few millimetres over a range of about 6000 km. The twin GRACE (Gravity Recovery and Climate Experiment) satellites were launched in 2002 by NASA and DLR (the German Aerospace Center). They are 200 - 250 km apart, in a near-polar orbit at an altitude of about 480 km. The GRACE space mission has allowed extremely accurate determinations of the Earth's gravitational field

Table 1: Main characteristics and orbital parameters of the satellites used in the LARES experiment.

| | LARES | LAGEOS | LAGEOS 2 | GRACE |
|---|---|---|---|---|
| Semimajor axis [km] | 7821 | 12270 | 12163 | 6856 |
| Eccentricity | 0.0008 | 0.0045 | 0.0135 | 0.005 |
| Inclination | 69.5° | 109.84° | 52.64° | 89° |
| Launch date | 13 Feb, 2012 | 4 May, 1976 | 22 Oct, 1992 | 17 Mar, 2002 |
| Mass [kg] | 386.8 | 406.965 | 405.38 | 432 |
| Number of CCRs | 92 | 426 | 426 | 4 |
| Diametre [cm] | 36.4 | 60 | 60 | |

and its temporal variations. For the main characteristics and orbital parameters of LARES, LAGEOS, LAGEOS 2 and GRACE see Table 1. The test of frame-dragging with the LAGEOS satellites was obtained by using the two observables quantities given by the two nodal rates of LAGEOS and LAGEOS 2 for the two main unknowns: the frame-dragging effect and the uncertainty in the Earth's quadrupole moment, $J_2$ [36]. If the Earth's gravitational potential is expanded in spherical harmonics, the even zonal harmonics are those of even degree and zero order. They represent the deviations from spherical symmetry of the gravitational potential of a body which are axially symmetric and which are also symmetric with respect to the equatorial plane of the body. The main secular drifts of the nodal longitude of a satellite are due to the Earth's even zonal harmonics. In particular, the largest node shift is by far due to the even zonal of degree two, $J_2$, i.e. the Earth's quadrupole moment [25]. To measure frame-dragging we either need to perfectly determine the Earth's even zonal harmonics or devise a method to neutralize the propagation of their uncertainties in our measurement.

## 3 LARES

LARES is a satellite of the Italian Space Agency (ASI) launched by the European Space Agency with the new launch vehicle VEGA (ESA-ASI-ELV-AVIO). It is a passive, spherical laser-ranged satellite (see Table 1). The LARES satellite was designed to approach as closely as possible an ideal test particle [37]. This goal was mainly achieved by adopting the following design requirements: (i) minimize the surface-to-mass ratio, (ii) reduce the number of parts, (iii) avoid any protruding component, (iv) use a non magnetic material and (v) avoid the painting of the satellite surface. The first requirement was implemented by using a tungsten alloy [38], the most dense material on Earth with an acceptable cost and good manufacturing characteristics. With a diameter of 36.4 cm and a total mass of 386.8 kg, the final mean density of the satellite is 15317 $kg/m^3$

which makes LARES the known orbiting object in the solar system with the highest mean density and the satellite with the lowest surface-to-mass ratio. The second requirement was achieved by building the satellite body out of one single piece of tungsten alloy, thus reducing thermal contact conductance and consequently the onset of thermal gradients. Temperature differences on the surface of the LAGEOS satellites produce in fact a tiny but not negligible perturbation: the thermal thrust [39]. To comply with the third requirement, the satellite interface with the separation system was limited only to four hemispherical cavities machined on the equator of the satellite. The fourth and fifth requirements were simply fulfilled by choosing a non magnetic tungsten alloy, although with slightly lower density than a magnetic tungsten alloy, with a proper surface treatment and with no painting [40].

## 4 Test of frame-dragging using LARES and the two LAGEOS satellites

The basic idea of the LARES space mission is to couple its orbital data with those of the two LAGEOS satellites in order to have three observable quantities provided by the nodal rates of the three satellites [41]. The three observables can then be used to determine the three unknowns: frame-dragging and the two uncertainties in the two lowest degree even zonal harmonics, $J_2$, and $J_4$ (i.e. the spherical harmonics of degree 2 and 4 and order 0). In such a way the two largest sources of uncertainty in the nodal drift are eliminated, providing an accurate measurement of frame-dragging within our systematic uncertainty of a few percent.

Here we report on our orbital analysis of the laser-ranging data of the LARES, LAGEOS and LAGEOS 2 satellites from 26 February 2012 until 6 September 2015 using a prominent state-of-the-art Earth's gravity field model, the GGM05S [42]. GGM05S is an Earth's gravity model released in 2013, based on approximately 10 years of GRACE data. It describes the Earth's spherical harmonics up to degree 180. The laser-ranging data of LARES, LAGEOS and LAGEOS 2 were collected from more than 30 ILRS stations all over the world (see Figure 1). We processed approximately one million normal points of LARES, LAGEOS and LAGEOS 2, corresponding to about 100 millions of laser ranging observations. The laser-ranging normal points were processed using NASA's orbital analysis and data reduction software GEODYN II [43], including the Earth's gravity model GGM05S, Earth's tides, solar radiation pressure, Earth's albedo, thermal thrust, Lunar, solar and planetary perturbations and Earth's rotation from Global Navigation Satellite System (GNSS) and Very Long Baseline Interferometry (VLBI).

The orbital residuals of a satellite are obtained by subtracting the observed orbital elements of the satellite with the computed ones. They provide a measurement of the orbital perturbations that, in the data reduction, are not included (un-modelled) or are modelled with some errors (mis-modelled) [26]. In particular, the residuals of the satellite's node are due to the errors in the Earth's even zonal harmonics and to the Lense-Thirring effect which we have not included in GEODYN II's modelling. The Lense-Thirring nodal shift, theoretically predicted by General Relativity, is about 30.7

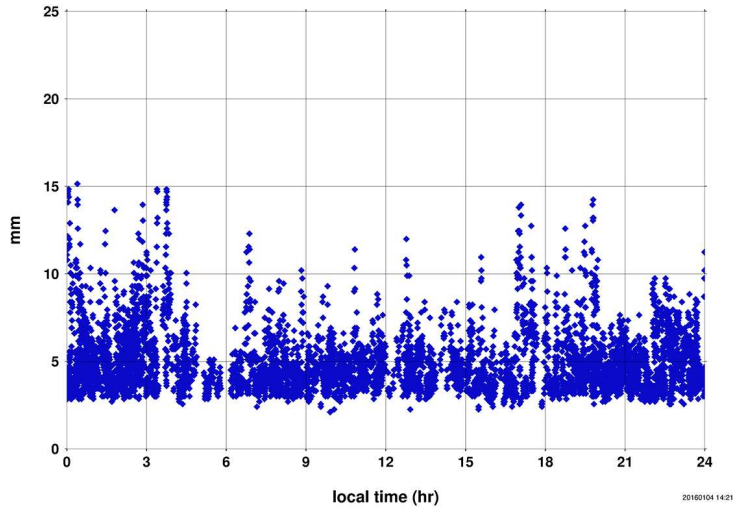



Figure 1: Root Mean Square (RMS) of the LARES normal points obtained from the laser-ranging observations of the Graz station of the ILRS during 2015. The average RMS of the LARES normal points is 4.83 millimeters (Courtesy of the ILRS [35]).

milliarcsec/yr on LAGEOS, about 31.5 milliarcsec/yr on LAGEOS 2 and about 118.4 milliarcsec/yr on LARES, the latter corresponding at the altitude of LARES to about 4.5m/yr.

Using the three observables provided by the three nodal rates of LAGEOS, LAGEOS 2 and LARES, we were able to eliminate not only the uncertainties in their nodal rates due to the errors in the even zonal harmonics $J_2$ and $J_4$ of the GGM05S model but also the uncertainties in their nodal rates due to the long and medium period tides contributing to the harmonics $J_2$ and $J_4$.

We fitted for the six largest tidal signals of LAGEOS, LAGEOS 2 and LARES, and for a secular trend, which produced:

$$\mu = (0.994 \pm 0.002) \pm 0.05 \tag{1}$$

Here $\mu = 1$ is the value of frame-dragging normalized to its GR value, 0.002 is the formal 1-sigma error (the postfit residuals of Figure 2 show a normal–Gaussian– distribution to good approximation) and 0.05 is the estimated systematic error due to the uncertainties in the Earth's gravity field model GGM05S and to the other error souces. We discuss systematic errors below.

In figure 3, we display the least squares secular trend fit of the combined residuals of LAGEOS, LAGEOS 2 and LARES prior to fitting for the tides. In contrast, in Figure 4 we show the secular trend obtained when including the six known periodical terms corresponding to the largest tidal signals observed on the satellite's nodes. The fit is obviously much tighter. These tidal signals were identified both by a Fourier analysis of

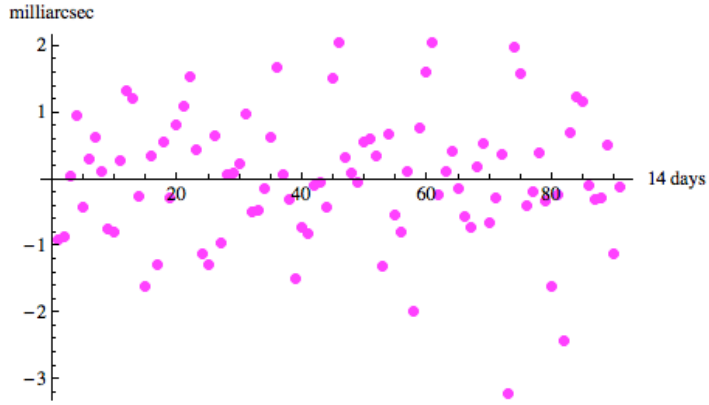



Figure 2: Combined residuals of LARES, LAGEOS and LAGEOS 2, over about 3.5 years of orbital observations, after the removal of six tidal signals and a constant trend.

the observed residuals and by analytical computations of the main tidal perturbations of the nodes of the satellites. Some of the signals observed in the nodal residuals correspond to the perturbations due to the main nongravitational perturbations.

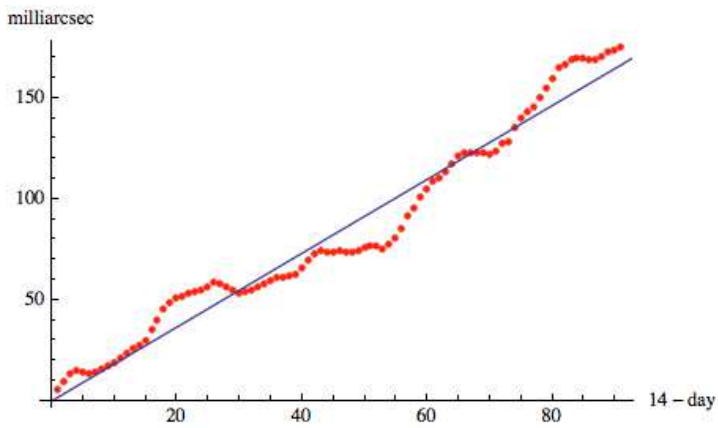



Figure 3: Fit of the combined orbital residuals of LARES, LAGEOS and LAGEOS 2 with a linear regression only.

The systematic errors in our measurement of frame-dragging with LARES, LAGEOS and LAGEOS 2 are mainly due to the errors in the even zonal harmonics of GGM05S, used in our orbital fits with GEODYN II, with degree strictly larger than four. To evaluate these systematic errors, we tripled the published calibrated errors (i.e. including

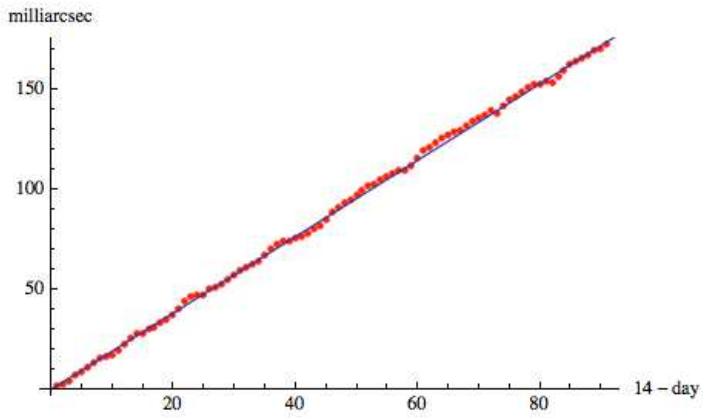



Figure 4: Fit of the combined orbital residuals of LARES, LAGEOS and LAGEOS 2 with a linear regression plus six periodical terms corresponding to six main tidal perturbations observed in the orbital residuals.

both the statistical and the systematic errors) of each even zonal coefficent of GGM05S (to multiply by a factor two or three is a standard technique in space geodesy to place an upper bound to the real error in the Earth's spherical harmonics) and then propagated these tripled errors into the nodes of LARES, LAGEOS and LAGEOS 2. We then found a systematic error of about 4% in our measurement of frame-dragging due to the Earth's even zonals.

Other smaller systematic errors are due to those long and medium period tides and non-gravitational perturbations either mis-modelled, or un-modelled. However, in our analysis we included the main tidal and non-gravitational perturbations, such as the direct radiation pressure from the Sun and the Earth, i.e. the albedo. Furthermore, the systematic errors due to the un-modelled or mis-modelled tidal and non-gravitational perturbations are periodical and their residual effect is quite small as clearly shown in the Fourier analysis of the post-fit orbital residuals shown in Figure 2. Previous error analyses [45, 46, 47, 48, 49, 29, 50, 51, 53, 54, 55] have confirmed that the systematic error due to tides, non-gravitatioanl perturbations and other error sources is at the level of approximately 3% and therefore the total Root Sum Squared (RSS) systematic error, including the systematic error due to the Earth's even zonals, is approximately at the level of 5% if the LARES, LAGEOS and LAGEOS 2 observations are used together with the Earth's gravity field model GGM05S.

Although we are quite pleased with the analysis to date of frame dragging including LARES, LAGEOS and LAGEOS 2, we consider this result only intermediate to a final determination. Our final result will present a careful restudy of systematics. We have been conservative here in quoting a 5% estimate of our systematic error. Extending the observation time of LARES and the other satellites will improve our understanding

of tidal contributions and will reduce the systematic error from that source. Different Earth's models lead to slightly different results, as is also the case for different orbital solvers. Completing a suite of solutions with different (up to date) Earth's models and different solvers will provide another estimate of the systematics. All these questions will be addressed in a forthcoming analysis of the measurement of frame-dragging using LARES, LAGEOS, LAGEOS 2 and GRACE.

However, we must also point out that the satellites LAGEOS, LAGEOS 2, and LARES will have tens of thousands of years on orbit, and will remain useful to laser-ranged science for an extremely long time. Eventually the retroreflectors may become degraded, but LAGEOS has shown no sign of this in its 40 years on orbit. Other laser-ranged satellites will be launched to join the current ones. All these satellites will be available while at the same time better Earth's models, better orbital solvers, and better models of nongravitational forces become available. The strength of this approach and these satellites is that they are available for innovative improvements in technique into the future.

## 5 Conclusions

Using the laser-ranged satellites LARES, LAGEOS and LAGEOS 2, and the Earth's gravity field described by the GGM05S model using the GRACE observations, we obtained a test of frame-dragging: $\mu = (0.994 \pm 0.002) \pm 0.05$, where $\mu = 1$ is the theoretical prediction of General Relativity, 0.002 is the 1-sigma statistical error and 0.05 is the estimated systematic error due to the uncertainties in the Earth's gravity field model GGM05S and to the other error sources.

## 6 Acknowledgements

We gratefully acknowledge the support of the Italian Space Agency, grants I/034/12/0, I/034/12/1 and 2015-021-R.0 and the International Laser Ranging Service for providing high-quality laser ranging tracking of the LARES satellites. E.C. Pavlis acknowledges the support of NASA Grants NNX09AU86G and NNX14AN50G. R. Matzner acknowledges NASA Grant NNX09AU86G and J.C. Ries the support of NASA Contract NNG06DA07C.

## References

[1] A. Einstein, "Die formale Grundlage der allgemeinen Relativitätstheorie", *Sitzungsber. Preuss Akad. Wiss.*, 1030, (1914).

[2] A. Einstein, "Herr Einstein sprach über den Grundgedanken der allgemeinen Relativitätstheorie und Anwendung dieser Theorie in der Astronomie", *Sitzungsher, Preuss Akad. Wiss.*, 315, (1915) ; "Zur allgemeinen Relativitätstheorie", *Preuss Akad. Wiss.*, 778–799, (1915); "Erklärung der Perihelbewegung des Merkur aus der allgemeinen

Relativitätstheorie", *Preuss Akad. Wiss.*, 831, (1915) ; "Die Feldgleichungen der Gravitation", *Preuss Akad. Wiss.*, 844, (1915).

[3] See also: D. Hilbert, "Die Grundlagen der Physik", Konigl. Gesell. d. Wiss. Göttingen, Nachr. Math.-Phys. Kl., 395 (1915).

[4] A. Einstein, "Die Grundlage der allgemeinen Relativitätstheorie", *Ann. Phys. (Leipzig)*, **49**, 769, (1916).

[5] Ya. B. Zeldovich, and I. D. Novikov, *Relativistic Astrophysics Vol. I Stars and Relativity* (Univ. Chicago Press, Chicago, 1971).

[6] S. Weinberg, *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity* (Wiley, New York, 1972).

[7] C.W. Misner, K.S. Thorne, and J.A. Wheeler, *Gravitation* Freeman, San Francisco, (1973).

[8] I. Ciufolini and J.A. Wheeler, *Gravitation and Inertia*, Princeton Univ. Press, (1995).

[9] S. Turyshev, Experimental Tests of General Relativity: Recent Progress and Future Directions, *Physics-Uspekhi*, **52**, 1-27 (2009).

[10] C. M. Will, The Confrontation between General Relativity and Experiment, *Living Rev. Relativity*, **17**, 4, (2014).

[11] B. P. Abbott et al., Observation of Gravitational Waves from a Binary Black Hole Merger, *Phys Rev. Lett.*, **116** 061102 (2016).

[12] S. Kopeikin, M. Efroimsky and G. Kaplan, *Relativistic Celestial Mechanics of the Solar System* (Wiley Online Library, 2011).

[13] N. Ashby, GNSS and Other Applications of General Relativity, in: "General Relativity: The most beautiful of Theories, Applications and trends after 100 years", Centennial Jubilee Volume of General Relativity, edited by C. Rovelli (De Gruyter, Berlin), 165-188, (2015).

[14] R. Penrose, Gravitational Collapse and Space-Time Singularities, *Phys. Rev. Lett.* **14**, 57-59, (1965).

[15] A. Riess et al., Observational evidence from supernovae for an accelerating universe and a cosmological constant, *Astron. J.*, **116**, 1009-1038, (1998).

[16] S. Perlmutter et al., Measurements of $\Omega$ and $\Lambda$ from 42 High-Redshift Supernovae, *Astrophys. J.*, **517**, 565-586, (1999).

[17] S. Perlmutter, Supernovae, Dark Energy, and the Accelerating Universe, *Phys. Today* **56**, 53-60, (2003).

[18] Robert R. Caldwell, "Dark Energy" *Physics World*, 37-42, (2004).

[19] I . H. Stairs, Testing General Relativity with Pulsar Timing, *Living Rev. Relativity*, **6**, (2003).

[20] R.F. O'Connell, Gravito-Magnetism in one-body and two-body systems: Theory and Experiments, Proc. of the International School of Physics "Enrico Fermi", Course CLXVIII, Atom Optics and Space Physics, Pages 149-159.

[21] C. Feiler, M. Buser, E. Kajari, W.P. Schleich, E.M. Rasel and R.F. O'Connell, New Frontiers at the Interface of General Relativity and Quantum Optics, *Space Sci Rev*, **148** 123147, (2009).

[22] I. Ciufolini, Dragging of Inertial Frames, *Nature*, **449**, 41-47, (2007).

[23] K.S. Thorne, R.H. Price and D.A. Macdonald, *The Membrane Paradigm*, Yale Univ. Press, NewHaven, (1986).

[24] J. Lense and H. Thirring, Uber den Einfluss der Eigenrotation der Zentralkörper auf die Bewegung der Planeten und Monde nach der Einsteinschen Gravitationstheorie, *Phys. Z.* , **19**, 156-163, (1918). See also English translation by B Mashhoon, F W Hehl, D S Theiss *Gen. Relativ. Gravit.*, **16**, 711-750, (1984).

[25] W.M. Kaula, *Theory of Satellite Geodesy*, Blaisdell, Waltham, (1966).

[26] I. Ciufolini, E.C. Pavlis, F. Chieppa, E. Fernandes-Vieira, and J. Perez-Mercader, Test of general relativity and measurement of the Lense-Thirring effect with two Earth satellites, *Science*, **279**, 2100-2103, (1998).

[27] S.C. Cohen and P.J. Dunn, (Eds.), LAGEOS Scientific Results, *J. Geophys. Res.*, **90** (B11), 9215, (1985).

[28] I. Ciufolini and E.C. Pavlis, A confirmation of the general relativistic prediction of the Lense-Thirring effect, *Nature*, **431**, 958-960, (2004).

[29] I. Ciufolini, E.C. Pavlis and R. Peron, Determination of frame-dragging using Earth gravity models from CHAMP and GRACE, *New Astronomy*, **11**, 527-550, (2006).

[30] I. Ciufolini, E.C. Pavlis,J. Ries, R. Koenig, G. Sindoni, A. Paolozzi, H. Neumayer, Gravitomagnetism and Its Measurement with Laser Ranging to the LAGEOS Satellites and GRACE Earth Gravity Models, In: General Relativity and John Archibald Wheeler, **367**, 371434. SpringerVerlag GmbH, Berlino DEU, (2010).

[31] I. Ciufolini, A. Paolozzi, E.C. Pavlis, J. Ries, V. Gurzadyan, R. Koenig, R. Matzner, R. Penrose and G. Sindoni, Testing General Relativity and gravitational physics using the LARES satellite *The European Physical Journal Plus*, **127**, 1-7, (2012).

[32] Ch. Reigber, F. Flechtner, R. Koenig, U. Meyer, K. Neumayer, R. Schmidt, P. Schwintzer and S. Zhu, GRACE Orbit and Gravity Field Recovery at GFZ Potsdam - First Experiences and Perspectives, *Eos. Trans. AGU*, **83**(47), Fall Meet. Suppl., Abstract G12B-03 (2002).

[33] B.D. Tapley, The GRACE Mission: Status and Performance Assessment, *Eos. Trans. AGU*, **83**(47), Fall Meet. Suppl., Abstract G12B-01 (2002).

[34] C.W. Everitt et al., 2011 Gravity Probe B: Final Results of a Space Experiment to Test General Relativity, *Phys. Rev. Lett*, **106**, 22110, (2011).

[35] M.R. Pearlman, J.J. Degnan and J.M. Bosworth, The International Laser Ranging Service, *Advances in Space Research*, **30**, 135-143.

[36] I. Ciufolini, On a new method to measure the gravitomagnetic field using two orbiting satellites, *Nuovo Cimento A*, **109**, 1709-1720, (1996).

[37] A. Paolozzi, I. Ciufolini, C. Paris, G. Sindoni. LARES: A new satellite specifically designed for testing general relativity, *International Journal of Aerospace Engineering*, Article ID 341384, 9 pages, (2015).

[38] A. Paolozzi, I. Ciufolini, F. Felli, A. Brotzu, D. Pilone. Issues on LARES Satellite Material, IAC-09.C2.4. Proceedings of the 60th International Astronautical Congress 2009 (IAC 2009), Daejeon, Korea, Oct. 12-16, 20095 (2009).

[39] I. Ciufolini, A. Paolozzi, C. Paris, G. Sindoni. The LARES satellite and its minimization of the thermal forces, In Metrology for Aerospace (MetroAeroSpace), 2014 IEEE, 299-303, 29-30 May 2014.

[40] A. Paolozzi, I. Ciufolini, C. Vendittozzi, F. Felli. Material and surface properties of LARES satellite, In: 63rd International Astronautical Congress (IAC 2012), Naples, Italy, 1-5 October 2012.

[41] I. Ciufolini, A. Paolozzi, E.C. Pavlis, J. Ries, R. Koenig, R. Matzner and G. Sindoni, The LARES Space Experiment: LARES Orbit, Error Analysis and Satellite Structure, in *John Archibald Wheleer and General Relativity* I. Ciufolini and R. Matzner eds., 371-434, Springer Verlag, (2010).

[42] B.D. Tapley, F. Flechtner, S. V. Bettadpur, M. M. Watkins, The status and future prospect for GRACE after the first decade, Eos Trans., Fall Meet. Suppl., Abstract G22A-01, (2013).

[43] D.E. Pavlis et al., GEODYN operations manuals (Contractor Report, Raytheon, ITSS, Landover MD, 1998).

[44] I. Ciufolini, F. Chieppa, D. Lucchesi and F. Vespe, Test of Lense- Thirring orbital shift due to spin, *Classical and Quantum Gravity*, 14:27012726, (1997).

[45] I. Ciufolini, A comprehensive introduction to the Lageos gravitomagnetic experiment: from the importance of the gravitomagnetic field in physics to preliminary error analysis and error budget, *Int. J. Mod. Phys. A*, **4**, 3083-3145, (1989).

[46] J.C. Ries, *Simulation of an experiment to measure the Lense-Thirring precession using a second LAGEOS satellite*, Ph. Dissertation, (Univ. of Texas, Austin, 1989).

[47] D.P. Rubincam, Drag on the LAGEOS satellite. *J. Geophys. Res.*, **95** (B11), 4881-4886, (1990).

[48] G.E. Peterson, *Estimation of the Lense-Thirring Precession Using Laser-Ranged Satellites*, Ph. Dissertation (Univ. of Texas, Austin, 1997).

[49] D.M. Lucchesi, Reassessment of the error modelling of non–gravitational perturbations on LAGEOS 2 and their impact in the Lense–Thirring determination. Part I, *Planet. Space Sci.*, **49**, 447-463, (2001).

[50] J.C. Ries, R.J. Eanes and M.M. Watkins, Confirming the Frame-Dragging Effect with Satellite Laser Ranging, *16th International Workshop on Laser Ranging*, 13-17 October 2008, Poznan, Poland.

[51] J.C. Ries, Relativity in Satellite Laser Ranging, American Astronomical Society, IAU Symposium 261. Relativity in Fundamental Astronomy: Dynamics, Reference Frames, and Data Analysis Virginia Beach, VA, USA, 27 April - 1 May (2009).

[52] I. Ciufolini, A. Paolozzi, E.C. Pavlis,J. Ries, R. Koenig, R. Matzner, G. Sindoni and H. Neumayer, Testing gravitational physics with satellite laser ranging, *The European Physical Journal Plus*, **126**, 119, (2011).

[53] J. Ries, I. Ciufolini, E. Pavlis, A. Paolozzi, R. Koenig, R. Matzner, G. Sindoni, and H. Neumayer, The Earth frame-dragging via laser-ranged satellites: a response to "some considerations on the present-day results for the detection of frame-dragging after the final outcome of GP-B" by Iorio L. *Europhysics Letters*, **96**, 1-5 (2011).

[54] V.G. Gurzadyan, I. Ciufolini, S. Sargsyan, G. Yegorian, S. Mirzoyan and A. Paolozzi, Satellite probing General Relativity and its extensions and Kolmogorov Analysis. *Europhysics Letters*, **102** 60002, 1-4 (2013).

[55] I. Ciufolini, B. Moreno Monge, A. Paolozzi, R. Koenig, G. Sindoni and G. Michalak, Monte Carlo Simulations of the LARES space experiment to test General Relativity and fundamental physics, *Class. Quantum Grav.* **30**, 235009, (2013).

# LARES Satellite Thermal Forces and a Test of General Relativity

Richard Matzner*, Phuc Nguyen*, Jason Brooks*, Ignazio Ciufolini†‡, Antonio Paolozzi‡§, Erricos C. Pavlis¶, Rolf Koenig‖, John Ries**, Vahe Gurzadyan††, Roger Penrose‡‡, Giampiero Sindoni§, Claudio Paris‡§, Harutyun Khachatryan†† and Sergey Mirzoyan††

*Theory Group, University of Texas at Austin, USA; Email: matzner2@physics.utexas.edu

† Dip. Ingegneria dell'Innovazione, Università del Salento, Lecce, Italy

‡ Museo Storico della Fisica e Centro Studi e Ricerche, Rome, Italy

§ Scuola di Ingegneria Aerospaziale, Sapienza Università di Roma, Italy

¶ Joint Center for Earth Systems Technology, (JCET), University of Maryland, Baltimore County, USA

‖ Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences, Potsdam, Germany

** Center for Space Research, University of Texas at Austin, USA

†† Center for Cosmology and Astrophysics, Alikhanian National Laboratory and Yerevan State University, Yerevan, Armenia

‡‡ Mathematical Institute, University of Oxford, UK

arXiv:1607.08787v2 [gr-qc] 23 Sep 2016

***Abstract*—We summarize a laser-ranged satellite test of frame dragging, a prediction of General Relativity, and then concentrate on the estimate of thermal thrust, an important perturbation affecting the accuracy of the test. The frame dragging study analysed $3.5$ years of data from the LARES satellite and a longer period of time for the two LAGEOS satellites. Using the gravity field GGM05S obtained via the Grace mission, which measures the Earth's gravitational field, the prediction of General Relativity is confirmed with a 1- $\sigma$ formal error of $0.002$, and a systematic error of $0.05$. The result for the value of the frame dragging around the Earth is $\mu = 0.994$, compared to $\mu = 1$ predicted by General Relativity. The thermal force model assumes heat flow from the sun (visual) and from Earth (IR) to the satellite core and to the fused silica reflectors on the satellite, and reradiation into space. For a roughly current epoch (days $1460 - 1580$ after launch) we calculate an average along track drag of $-0.50 pm/sec^2$.**

## I. Dragging of Inertial Frames

In 1918, Lense and Thirring [1] described the frame-dragging of an orbiting body, in the limiting case valid around Earth of slow rotation of the central body and weak gravitational field: $\dot{\mathbf{\Omega}} = 2\mathbf{J}/(a^3\,(1-e^2)^{3/2})$. Here $\mathbf{\Omega}$, $a$ and $e$ are the longitude of the *ascending node*, the semimajor axis and the eccentricity of the orbiting body, while $\mathbf{J}$ is the angular momentum of the central body. We recall that the orbital plane intersects the equator in two nodes, one of which, the ascending node, is crossed by the orbiting body from southern to northern hemisphere [2]. The Earth satisfies the conditions of "weak gravitational field" and "slowly rotating", so this formula describes the behavior of Earth satellites. This *Eastward* precession has been observed and measured using the two LAGEOS satellites [3]–[7] and a model of the gravitational field of the Earth provided by the GRACE mission [8], [9], with improving errors to about 10%. (Satellite Laser Ranging allows range measurement with an accuracy that can reach a few millimeters [10].) The LAGEOS tests of frame-dragging were used to set limits on some string theories equivalent to Chern-Simons gravity [11]. Another satellite experiment, *Gravity Probe B* (GPB), was put into orbit in 2004 and verified frame dragging (via a related gyroscope precession effect) with approximately 20% accuracy [12]. Very recently, preliminary analysis of the LARES laser ranged satellite along with the two LAGEOS satellites and latest GRACE results, produced a frame dragging consistent with that predicted by GR, with estimated 5% errors [13]. We summarize the results of that analysis below and then go on to model thermally induced forces on LARES. LARES is a satellite with a spherical tungsten alloy core which carries 92 *cube corner reflectors* (CCRs) for laser ranging.

## II. LARES frame-dragging mission

The node shift of an orbiting body is dominated by the classical Newtonian effect due to the axially symmetric components of deviations of the gravity field of the central body from sphericity. These deviations are specifically due to the *even zonal harmonics*. Their effect decreases with increasing order, so the main perturbing effect is due to the even zonal of degree two, $J_2$, (Earth's quadrupole moment) [2]. The Newtonian effects precess the orbital plane of LARES westward at a rate of $\approx 1.7$ degrees/day [14]!

The LARES experiment combines data from the three satellites (Table I) LARES (Italian Space Agency - ASI), LAGEOS (NASA) and LAGEOS 2 (NASA and ASI) to obtain three observable quantities provided by the nodal rates of the three satellites [15]. The three observables can then be used to determine three unknowns: frame-dragging and the two uncertainties in the two lowest degree even zonal harmonics, $J_2$, and $J_4$. We analyzed data provided by the International Laser Ranging Service (ILRS) for the

TABLE I
MAIN CHARACTERISTICS AND ORBITAL PARAMETERS OF THE SATELLITES OF THE LARES EXPERIMENT.

| | LARES | LAGEOS | LAGEOS 2 | GRACE |
|---|---|---|---|---|
| Semimajor axis [km] | 7821 | 12270 | 12163 | 6856 |
| Eccentricity | 0.0008 | 0.0045 | 0.0135 | 0.005 |
| Inclination | 69.5° | 109.84° | 52.64° | 89° |
| Launch date | 13 Feb, 2012 | 4 May, 1976 | 22 Oct, 1992 | 17 Mar, 2002 |
| Mass [kg] | 386.8 | 406.965 | 405.38 | 432 |
| Number of CCRs | 92 | 426 | 426 | 4 |
| Diametre [cm] | 36.4 | 60 | 60 | |

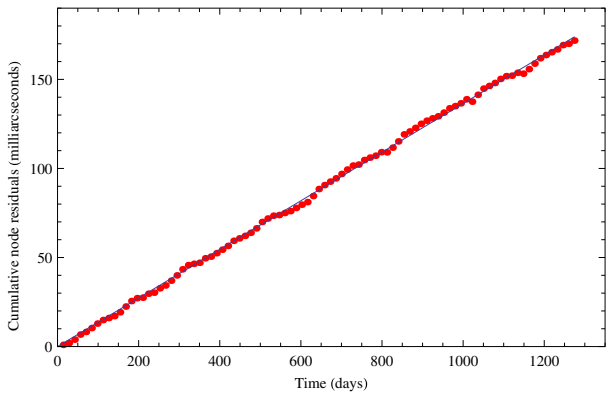



Fig. 1. Fit of the combined orbital residuals of LARES, LAGEOS and LAGEOS 2 with six periodical terms corresponding to six main tidal perturbations observed in the orbital residuals removed. The slope is $\mu = 0.994$ where $\mu = 1$ is the GR result [13].

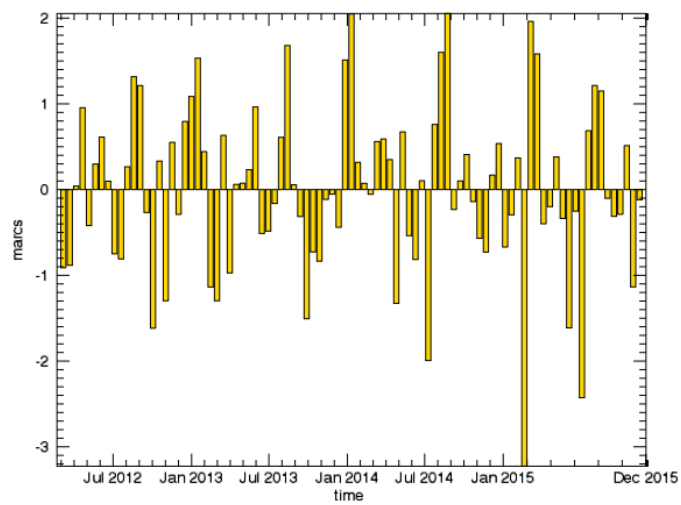



Fig. 2. Residuals after the removal of the six tidal signals and the constant trend [13].

three satellites from 26 February 2012 to 6 September 2015. The laser-ranging normal points were analysed using NASA's orbital determination program GEODYN II [16], including the Earth gravity model GGM05S, Earth tides, solar radiation pressure, Earth albedo, thermal thrust, lunar, solar and planetary perturbations and Earth rotation from Global Navigation Satellite Systems and Very Long Baseline Interferometry. GGM05S is a state-of-the-art Earth gravity model released in 2013, based on approximately 10 years of GRACE data. It describes the Earth's spherical harmonics up to degree 180 [17].

The orbital residuals of a satellite are obtained by subtracting the computed orbital elements of the satellite from the observed ones. The residuals of the satellite's node are due to the errors in the Earth's even zonal harmonics and to the Lense-Thirring effect which is not included in GEODYN II's modelling. General Relativity predicts frame-dragging of about 30.7, 31.5, and 118.4 milliarcsec/year on LAGEOS, LAGEOS 2 and LARES, respectively. At the LARES orbit this corresponds to a nodal displacement of about 4.5 meters/year.

To have a significantly accurate fit of the observed residuals it is necessary to include in the modeling the main periodic tidal signals that were singled out not only analytically but also by a Fourier analysis. The resulting fit, including the secular trend and six tides, produced (Fig. 1):

$$\mu = (0.994 \pm 0.002) \pm 0.05 \tag{1}$$

Here $\mu = 1$ is the value of frame-dragging normalized to its GR value, 0.002 is the formal 1-sigma error (the postfit residuals in Figure 2 show a normal –Gaussian– distribution to good approximation) and 0.05 is the estimated systematic error caused by uncertainties in GGM05S in the even zonal harmonics with degree greater than four, and other error sources. For a conservative estimate of the systematic errors we took three times the published values of the calibrated errors of the coefficients of the harmonic expansion of the Earth gravitational field. This is a common approach in space geodesy for Earth gravity field error estimation (sometimes

taking a factor of two instead of three). Those tripled errors were then propagated to the nodes of the three satellites of the LARES experiment providing a systematic error of about 0.04 in frame-dragging caused by the uncertainties in the knowledge of the Earth's gravity field even zonal harmonics. Previous error analyses [5], [18]–[27] have shown that the systematic errors associated with all the other perturbations, i.e. tides, non-gravitational perturbations and other unidentified effects, amount to about 0.03. The combination of all those errors, using the total Root Sum Squared (RSS), provides a value of 5%.[1] Note that the analysis described above and reported in [13] solves for thermally generated forces from the data, rather than predicting them from a model. The rest of this paper describes a model for those forces [28].

## III. Thermal thrust and temperature distribution on LARES' surface

Thermal thrust is the most subtle of the non-gravitational perturbations acting on LARES. It is small and difficult to model because is influenced by many parameters that affect the temperature distribution on the body. First the two sources: solar radiation and Earth infrared (IR) radiation, that are strongly dependent on the orbit and sun position, which affect the occurrence and duration of eclipses. As a consequence the heat exchange situation will change constantly thus making it necessary to evaluate the temperature distribution and the relevant thermal thrust day by day. Then we have the influence of the material body of the satellite and the way the parts are mounted. Because the metal sphere is a very good heat conductor, its temperature is quite uniform, so it radiates heat isotropically and produces little thermal thrust.

A different situation occurs for the CCRs, which are mounted loosely between plastic mounting rings. This mounting, designed to avoid thermal stresses on the CCRs, makes each CCR practically thermally isolated from the satellite body and from the other CCRs. In addition each CCR is made of Suprasil 311, a special grade glass that has a substantial heat capacity (though slightly lower than conventional optical glass) and significant absorption of infrared radiation. Thus during exposure to heat source it will heat with a delay and when in the shadow it will cool with a delay. Each of the CCRs can have a different temperature, so unequal radiation from CCRs can lead to an effective nonisotropic thermal radiation and an acceleration of the satellite.

Analysis of data of the first 4 months in orbit showed an anomalous along-track acceleration of LARES that on average was about $-0.4pm/s^2$ $(pm := picometer)$.[2] An earlier algorithm was able to closely match this observation by using parameter values for the glass IR absorbance $\alpha_{gl,IR} \equiv$ emissivity $\epsilon_{gl,IR}$ of the CCRs equal to 0.60 and visual absorbance $\alpha_{gl,vis}$ equal to 0.15 [29]. These are consistent with surface-contaminated glass on orbit, and with experiments carried out on orbit to determine $\alpha_{gl,vis}$ [30], [31].

## IV. Satellite structure

The CCRs are distributed uniformly on LARES surface along each parallel. They are disposed in each parallel as shown in the following table where $I$, $n_I$ and $\theta_I$ are the parallel number, the number of CCRs in the parallel and the colatitude of the parallel.

| Parallel | $n_I$ | $\theta_I(deg)$ | Parallel | $n_I$ | $\theta_I(deg)$ |
|---|---|---|---|---|---|
| I | 1 | 0 | -V | 16 | 100 |
| II | 5 | 20 | -IV | 14 | 120 |
| III | 10 | 40 | -III | 10 | 140 |
| IV | 14 | 60 | -II | 5 | 160 |
| V | 16 | 80 | -I | 1 | 180 |

A CCR is basically a corner of a cube. Its shape is completed with a cylindrical part that provides a circular front face of diameter of 1.5 inches. The CCR is loosely maintained in place by two plastic mounting rings (made of KEL-F) that engage three protruding tabs of the custom made CCRs. The mounting system is completed with a retainer ring and three screws that lock the plastic rings in place. Screws and retainer rings are made from the same alloy as the entire satellite. This design allows a stress free mounting for the CCRs but introduces uncertainties in the temperature estimation of the CCRs. We regard the tungsten alloy retainer rings as being at the same temperature as the satellite body [29].

## V. Orbit

The European Space Agency VEGA launcher, developed by ELV and Avio, lifted off with the LARES satellite on-board on February 13, 2012 at 10:00 UTC (07:00 local time) from Kourou spaceport in French Guyana [32]. In the following celestial coordinates are used.[3] The longitude of the ascending node on day $k$ is $\Omega(k) \approx (220 - 1.7k)\pi/180$, and this nodal drift is caused mainly by the Newtonian gravitational effect of $J_2$. Time $t$ spans between 0 and the orbital period $T$ where at time $t = 0$ the satellite is at the ascending node.

The experimental values of orientation and frequency of the spin of LARES are found the Ref. [33]. The orientation of the spin is at $RA = 12^h 22^m 48^s$ $(RMS = 49^m)$ and $Dec = -70.4$ degrees $(RMS = 5.2$ degrees). The spin axis is taken coincident with the south-north axis of the satellite. The eddy currents induced in the satellite body by the Earth's magnetic field cause a spin down torque

[1] A reduction of the systematic errors can be obtained by an increase of the number of years of observations (preferably more than 7 years from launch date). Furthermore robustness of the results can be increased by using different and updated Earth gravity field models and independent orbital determination programs for the orbital analysis.

[2] This remarkably small residual acceleration is smaller than that of any other satellite, including the LAGEOS satellite, whose mean along-track acceleration is about $-2pm/s^2$.

[3] The right-handed system aligned with the Earth's pole ($z$) and the vernal equinox ($x$).

that reduces the spin frequency of the satellite, as [33]: $\omega(k) = (0.546\mathrm{rad/s})\exp -0.00322509k$. This expression predicts the spin rate to be now ($\approx$ day 1500 in orbit) of order $\approx 5/orbit$, so CCR temperature equilibration rates are comparable to the satellite spin rate, and large individual-CCR effects may arise. We use direct numerical integration (built in functions in *Mathematica* [34]) of the coupled nonlinear heat-transfer equations for the 93 separate pieces of the satellite (the tungsten alloy core and 92 CCRs). Each piece is considered separately isothermal. Our analysis is close to Slabinski's [35] LAGEOS study. He modeled the thermal behaviour of LAGEOS and integrated the nonlinear heat balance equations with the aim of modeling the along track component of the thermal thrust of the satellite. In the present work all the components of the thermal forces are estimated as a function of time along the orbit. We also use a comparison Fourier code [29], which linearizes the equations and assumes the spin frequency is an integer multiple (here, exactly 5) of the orbital frequency. The Fourier code provides initial data for the nonlinear code, which uses the exact satellite spin rate. The nonlinear set is then run for several orbits (here 8) with fixed orbital parameters to relax out any effect from the Fourier provided initial conditions. The thermal force on the $ith$ CCR is perpendicular to the face (inward toward the center of the satellite) and with magnitude $F_i = 2\pi R^2 \epsilon_{gl,IR} \sigma T_i^4/(3c)$. The vector sum of the forces of all the CCRs produces the net force on LARES satellite The along track force is a unit vector along the orbit dotted into this vector sum. For details concerning the satellite design, the determination of the view factors between the CCRs and the tungsten alloy cavities housing the CCRs, the orbit geometry and relevant eclipses (Fig. 3) as well as the spin orientation, refer to Ref. [29].

In the current slow spin regime, Fourier methods produce valuable average values for CCR temperatures (which in that method are assumed to be equal in parallels), and for the daily along-track average drag. However there are significant deviations from the Fourier results during the orbit. The calculated average along track drag swings between $-0.70pm/s^2$ and $-0.25pm/s^2$ and averages $-0.50pm/s^2$ over days 1460 - 1580; Fig. 4. But the calculated instantaneous accelerations have excursions that are about an order of magnitude larger than the resulting orbit-averaged drag (Figs. 5 - 8). Figures 9 and 10 show the time dependence of the core temperature and of the individual CCRs in parallel $V$.

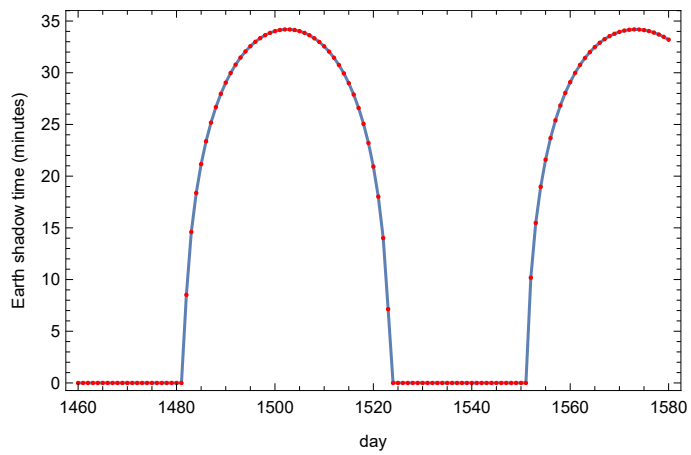



Fig. 3. The eclipse duration as a function of number of days since launch for days 1460 - 1580.

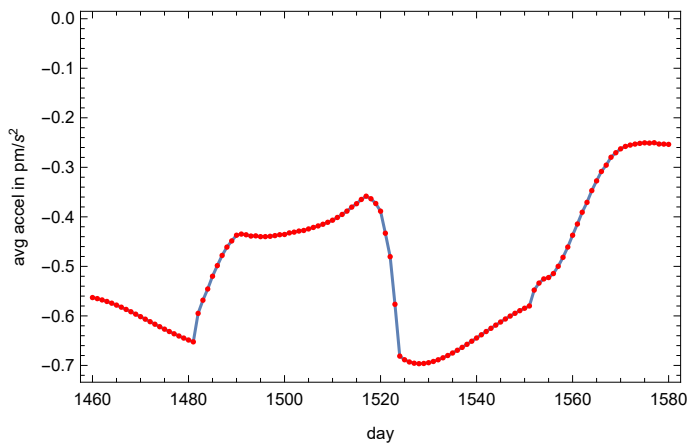



Fig. 4. Time series of the calculated daily average thermal drag; negative values indicate *drag*. The average calculated thermal drag (average of daily averages) over these 120 days is $-0.50pm/s^2$.

## Acknowledgment


We thank Victor Slabinski for very useful comments. This material is supported by the Texas Cosmology Center (TCC). TCC is supported by the College of Natural Sciences, the Department of Astronomy at the University of Texas at Austin, and the McDonald Observatory.

We gratefully acknowledge Italian Space Agency grants I/034/12/0, I/034/12/1 and 2015-021-R.0 and the International Laser Ranging Service for providing high-quality laser ranging tracking of the LARES satellites. E.C. Pavlis acknowledges the support of NASA Grants NNX09AU86G and NNX14AN50G. R. Matzner acknowledges NASA Grant NNX09AU86G and J.C. Ries NASA Contract NNG12VI01C.


## References


[1] J. Lense and H. Thirring, Uber den Einfluss der Eigenrotation der Zentralkörper auf die Bewegung der Planeten und Monde nach der Einsteinschen Gravitationstheorie, *Phys. Z.* , **19**, 156-163, (1918). See also English translation by B Mashhoon, F W Hehl, D S Theiss *Gen. Relativ. Gravit.*, **16**, 711-750, (1984).

[2] W.M. Kaula, *Theory of Satellite Geodesy*, Blaisdell, Waltham, (1966).

[3] S.C. Cohen and P.J. Dunn, (Eds.), LAGEOS Scientific Results, *J. Geophys. Res.*, **90** (B11), 9215, (1985).

[4] I. Ciufolini and E.C. Pavlis, A confirmation of the general relativistic prediction of the Lense-Thirring effect, *Nature*, **431**, 958-960, (2004).

[5] I. Ciufolini, E.C. Pavlis and R. Peron, Determination of frame-dragging using Earth gravity models from CHAMP and GRACE, *New Astronomy*, **11**, 527-550, (2006).

[6] I. Ciufolini, E.C. Pavlis,J. Ries, R. Koenig, G. Sindoni, A. Paolozzi, H. Neumayer, Gravitomagnetism and Its Measurement with Laser Ranging to the LAGEOS Satellites and GRACE Earth Gravity Models, In: General

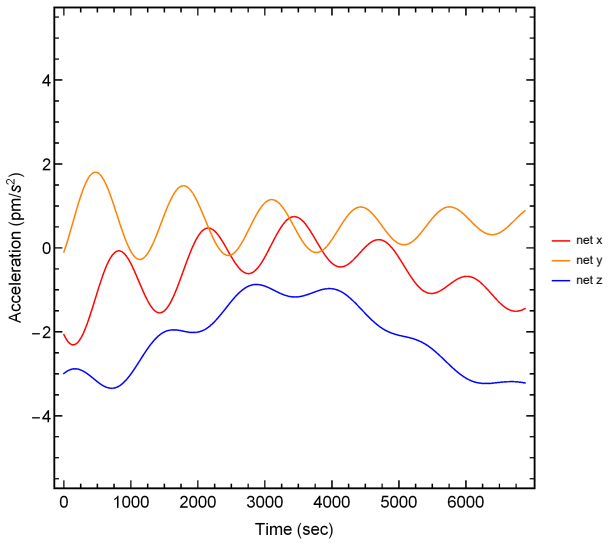



Fig. 5. Day 1470; there is no eclipse on this day. Calculated net acceleration $(pm/s^2)$ in components (in celestial coordinate frame). The spin rate is approximately $5/orbit$, and this introduces obvious variations into the accelerations.

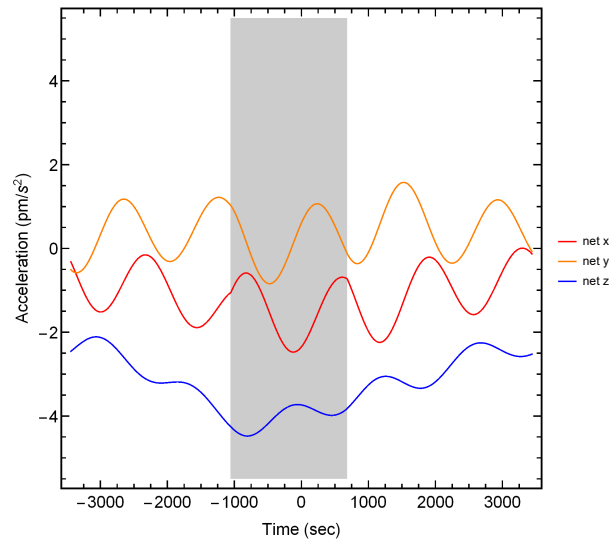



Fig. 7. Day 1490: Calculated net acceleration in components. An eclipse occurred between about $-1050$ and 700 seconds in the orbit on this day, and the satellite crossed the ascending node during the eclipse. We plot one full orbital period, with the node crossing in the center of the plot to more clearly show eclipse-induced effects. The onset and the end of the eclipse introduce noticeable features into time derivative of some of the force components and into the drag (Fig 8), and there are clearly visible $\approx 5/orbit$ spin effects.

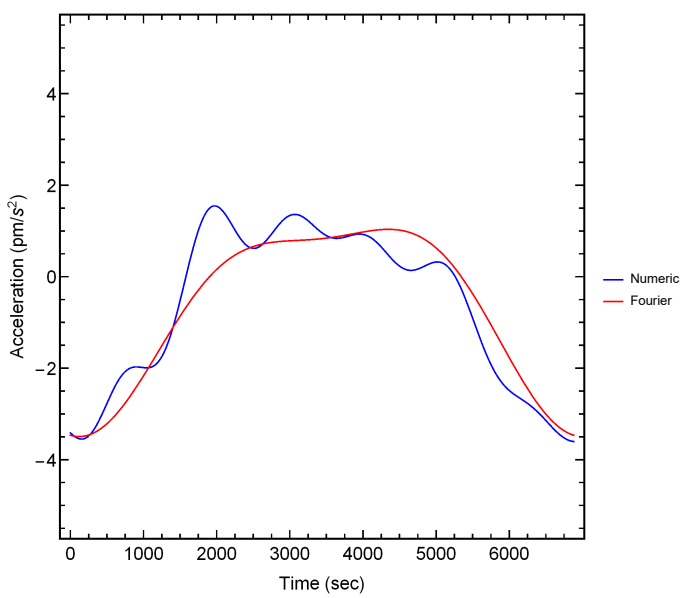



Fig. 6. Day 1470: Along-track acceleration in numerical simulation (blue curve). Negative values indicate *drag*. Comparison fast-spin (assuming all CCRs in a row have the same temperature; see [29]) Fourier computation (sum of the constant, first and second harmonics of the orbital frequency): smoother, red curve, over one orbital period. Note that amplitudes of the force components (Fig 5) and of the along-track acceleration are much larger (almost an order of magnitude larger) than the average drag for that day $(-0.60pm/s^2$; see Figure 4).

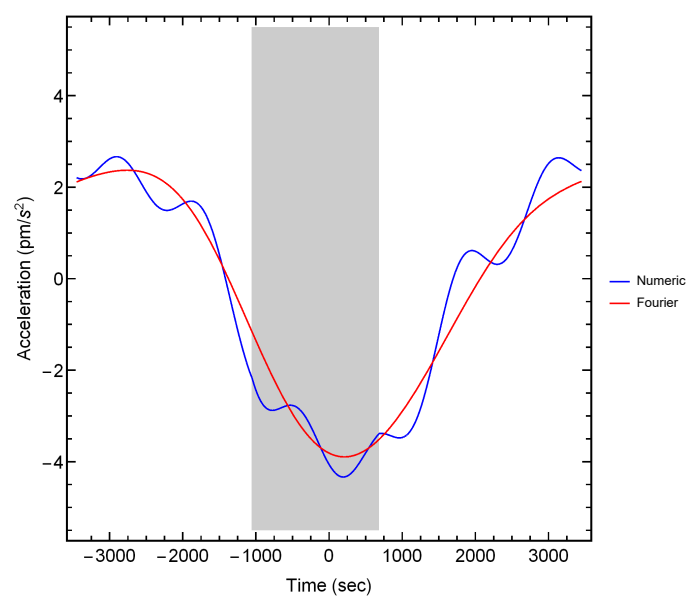



Fig. 8. Day 1490: Calculation of along-track acceleration in numerical simulation (blue curve); Comparison fast-spin Fourier computation (smoother, red curve). Average drag for that day :$-0.40pm/s^2$; see Figure 4.

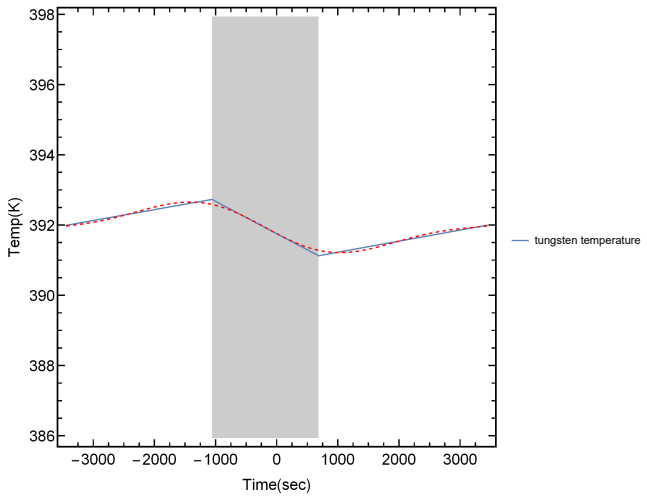



Fig. 9. Day 1490: Computed tungsten alloy core temperature. The dotted line is a Fourier result (constant plus first and second orbital harmonic). The eclipse introduces noticeable sudden changes in the time derivatives of the temperature of the metal and of the CCRs (Fig. (10)). The range of both the Figs. 9 and 10 is $12K$.

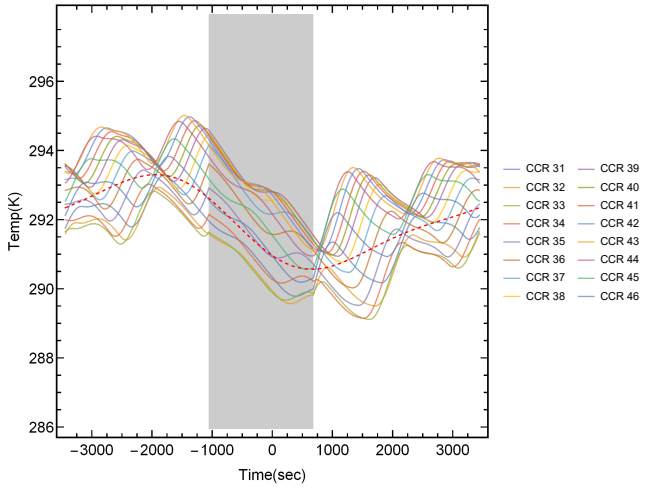



Fig. 10. Day 1490: Calculated individual CCR temperatures for parallel $V$ (see Section IV).


Relativity and John Archibald Wheeler, **367**, 371434. SpringerVerlag GmbH, Berlino DEU, (2010).

[7] I. Ciufolini, A. Paolozzi, E.C. Pavlis, J. Ries, V. Gurzadyan, R. Koenig, R. Matzner, R. Penrose and G. Sindoni, Testing General Relativity and gravitational physics using the LARES satellite *The European Physical Journal Plus*, **127**, 1-7, (2012).

[8] Ch. Reigber, F. Flechtner, R. Koenig, U. Meyer, K. Neumayer, R. Schmidt, P. Schwintzer and S. Zhu, GRACE Orbit and Gravity Field Recovery at GFZ Potsdam - First Experiences and Perspectives, *Eos. Trans. AGU*, **83**(47), Fall Meet. Suppl., Abstract G12B-03 (2002).

[9] B.D. Tapley, The GRACE Mission: Status and Performance Assessment, *Eos. Trans. AGU*, **83**(47), Fall Meet. Suppl., Abstract G12B-01 (2002).

[10] M.R. Pearlman, J.J. Degnan and J.M. Bosworth, The International Laser Ranging Service, *Advances in Space Research*, **30**, 135-143.

[11] T. L. Smith, A. Erickcek, R. Caldwell and M. Kamionkowski, Effects of ChernSimons gravity on bodies orbiting the Earth. *Physical Review* D 77 024015 (2008).

[12] C.W. Everitt et al., 2011 Gravity Probe B: Final Results of a Space Experiment to Test General Relativity, *Phys. Rev. Lett*, **106**, 22110, (2011).

[13] "A test of general relativity using the LARES and LAGEOS satellites and a GRACE Earth gravity model; Measurement of Earths dragging of inertial frames" (Ignazio Ciufolini, Antonio Paolozzi, Erricos C. Pavlis, Rolf Koenig, John Ries, Vahe Gurzadyan, Harutyun Khachatryan, Richard Matzner, Roger Penrose, Giampiero Sindoni, Claudio Paris, Sergey Mirzoyan) *European Physics Journal* **C76** 120; DOI: 10.1140/epjc/s10052-016-3961-8 (2016).

[14] I. Ciufolini, A. Paolozzi, E. Pavlis, J. Ries, R. Koenig, R. Matzner, V. Gurzadyan, G. Sindoni, C. Paris, H. Neumayer, *Initial Orbit Determination Results For The Lares Satellite*, [29]AC-12,B4,2,1,x15153, *63rd International Astronautical Congress, Naples, Italy* (2012).

[15] I. Ciufolini, A. Paolozzi, E.C. Pavlis, J. Ries, R. Koenig, R. Matzner and G. Sindoni, The LARES Space Experiment: LARES Orbit, Error Analysis and Satellite Structure, in *John Archibald Wheleer and General Relativity* I. Ciufolini and R. Matzner eds., 371-434, Springer Verlag, (2010).

[16] D.E. Pavlis et al., GEODYN operations manuals (Contractor Report, Raytheon, ITSS, Landover MD, 1998).

[17] B.D. Tapley, F. Flechtner, S. V. Bettadpur, M. M. Watkins, The status and future prospect for GRACE after the first decade, Eos Trans., Fall Meet. Suppl., Abstract G22A-01, (2013).

[18] I. Ciufolini, A comprehensive introduction to the LAGEOS gravitomagnetic experiment: from the importance of the gravitomagnetic field in physics to preliminary error analysis and error budget, *Int. J. Mod. Phys. A*, **4**, 3083-3145, (1989).

[19] J.C. Ries, *Simulation of an experiment to measure the Lense-Thirring precession using a second LAGEOS satellite*, Ph. Dissertation, (Univ. of Texas, Austin, 1989).

[20] D.P. Rubincam, Drag on the LAGEOS satellite. *J. Geophys. Res.*, **95** (B11), 4881-4886, (1990).

[21] G.E. Peterson, *Estimation of the Lense-Thirring Precession Using Laser-Ranged Satellites*, Ph. Dissertation (Univ. of Texas, Austin, 1997).

[22] D.M. Lucchesi, Reassessment of the error modelling of non–gravitational perturbations on LAGEOS 2 and their impact in the Lense–Thirring determination. Part I, *Planet. Space Sci.*, **49**, 447-463, (2001).

[23] J.C. Ries, R.J. Eanes and M.M. Watkins, Confirming the Frame-Dragging Effect with Satellite Laser Ranging, *16th International Workshop on Laser Ranging*, 13-17 October 2008, Poznan, Poland.

[24] J.C. Ries, Relativity in Satellite Laser Ranging, American Astronomical Society, IAU Symposium 261. Relativity in Fundamental Astronomy: Dynamics, Reference Frames, and Data Analysis Virginia Beach, VA, USA, 27 April - 1 May (2009).

[25] J. Ries, I. Ciufolini, E. Pavlis, A. Paolozzi, R. Koenig, R. Matzner, G. Sindoni, and H. Neumayer, The Earth frame-dragging via laser-ranged satellites: a response to "some considerations on the present-day results for the detection of frame-dragging after the final outcome of GP-B" by Iorio L. *Europhysics Letters*, **96**, 1-5 (2011).

[26] V.G. Gurzadyan, I. Ciufolini, S. Sargsyan, G. Yegorian, S. Mirzoyan and A. Paolozzi, Satellite probing General Relativity and its extensions and Kolmogorov Analysis. *Europhysics Letters*, **102** 60002, 1-4 (2013).

[27] I. Ciufolini, B. Moreno Monge, A. Paolozzi, R. Koenig, G. Sindoni and G. Michalak, Monte Carlo Simulations of the LARES space experiment to test General Relativity and fundamental physics, *Class. Quantum Grav.* **30**, 235009, (2013).

[28] Jason W. Brooks & Richard Matzner "Modelling LARES temperature distribution and thermal drag II: Numerical computation of current-epoch thermal forces" *European Physics Journal Plus* (submitted, 2016).

[29] Phuc H. Nguyen & Richard Matzner "Modelling LARES temperature distribution and thermal drag" *European Physics Journal Plus* **130** 206; DOI: 10.1140/epjp/i2015-15206-2 (2015).

[30] W.R. Pence, T.J. Grant, *AIAA-81-1186* $\alpha_s$ *Measurements of Thermal Control Coatings on Navstar Global Positioning System Spacecraft* (Rockwell International Corp., Downey, CA, 1981).

[31] N.L. Hyman, Comsat Techn. Rev. **11** 1 (1981).

[32] A. Paolozzi, I. Ciufolini, Acta Astron. **91**, 313 (2013).

[33] D. Kucharski, T. Otsubo, G. Kirchner, G. Bianco, Adv. Space Res. **50**, 1473 (2012).

[34] https://www.wolfram.com/mathematica/

[35] V.J. Slabinski, Celest. Mech. Dyn. Astron. **66**, 131 (1996).

# Correlated "noise" in LIGO gravitational wave signals: an implication of Conformal Cyclic Cosmology

by Roger Penrose
Mathematical Institute, Oxford, UK

*Abstract*

It has recently been reported by Cresswell *et al*. [1] that correlations in the *noise* surrounding the observed gravitational wave signals, GW150194, GW151226, and GW170194 were found by the two LIGO detectors in Hanford and Livingston with the same time delay as the signals themselves. This raised some issues about the statistical reliability of the signals themselves, which led to much discussion, the current view appearing to support the contention that there is something unexplained that may be of genuine astrophysical interest [2]. In this note, it is pointed out that a resolution of this puzzle may be found in a proposal very recently put forward by the author [3], see also [4], that what seems to be spuriously generated noise may in fact be gravitational events caused by the decay of dark-matter particles (erebons) of mass around $10^{-5}$g, the existence of such events being a clear implication of the cosmological scheme of conformal cyclic cosmology, or CCC [5], [6]. A brief outline of the salient points of CCC is provided here, especially with regard to its prediction of erebons and their impulsive gravitational signals.

## 1. Dark matter decay, as an astrophysical source of impulsive gravitational signals

A dispute has recently arisen concerning correlations that have been pointed out between "noise" events in the data collected by the LIGO gravitational wave detectors at Hanford WA and Livingston LA. According to Cresswell *et al* [1], the time delay (6.9ms) between the reception of the gravitational wave signal GW150509 at these two locations was also the same time delay that provided a surprising correlation between the noise in the signal, occurring before, after, and during the actual wave form of the gravitational-wave signal—where that wave form is convincingly interpreted as coming from an inspiralling pair of black holes. A similar effect was observed also in the black-hole encounter events GW151226 and GW170104. This correlation is, on the face of it, hard to account for, since the concept of "noise" here, would normally be expected to be some local effect, in the neighbourhood of each of the two detectors, at Hanford and, Livingston, which are separated by some 3000km. Such correlation would appear to imply that a significant component of this "noise" has actually a cosmological or astrophysical origin, rather than a local one. If such an external explanation can be provided, then this could resolve the dispute referred to at the beginning of this section.

In this note, I point out that there is a clear astrophysical candidate for what could well be a major component of these "noise" signals [3]. I contend that they could, in large part, be actual gravitational signals, of an unusual type, these being clearly predicted to occur on the basis of a particular cosmological proposal, referred to as *conformal cyclic cosmology*, or CCC, which was originally put forward in 2005/6. Supporting observational evidence for CCC was subsequently published [6], [7], [8], [9], though disputed by others [10]. The

unusual gravitational signals referred to above were recognized as being implications of that theory only very recently (see [3]). A brief outline of CCC, and its significant observational support will be given in Section 3. In this section, I simply assert the implications of CCC that are relevant to the issue at hand without providing justification from CCC or otherwise. Such justification will be provided in Section 4.

It is indeed a reasonably robust implication of CCC that dark matter consists of particles—that I refer to as *erebons* (after the ancient Greek God of Darkness *Erebos*) that have a *Planck mass*, or thereabouts (~$10^{-5}g$), and a decay lifetime that could well be in excess of some $10^{10}$ years. Erebons are scalar particles (having zero spin), and interact only gravitationally (this much being a necessary feature of their key role in CCC as a kind of scalar conformal partner to gravitation). Because of its large mass, each erebon should behave effectively like a *classical* particle—according to various proposals that "self-measurement" should occur in quantum systems that involve sufficient mass displacement [11], [12], [13], [14], [15]. Moreover, its decay would, correspondingly, take the form of, a total conversion into a highly oscillatory *classical* gravitational wave, travelling outwards from the erebon's location, in an effectively classical impulsive spherical shell.

Since the oscillation frequency would be (initially, at least) of the absurdly large order of a Planck frequency (~$10^{43}$ Hz), there is no way that the oscillatory nature of the wave would be picked up in any presently feasible gravitational wave detector. Instead, we have to consider that, in effect, the wave would behave more like an instantaneous impulse that would be of a confusing nature for those who are used to studying ordinary macroscopic gravitational waves. We think of such a macroscopic gravitational wave as having the effect, on any material that it encounters, as inflicting a *squashing* (inward acceleration) in one direction transverse to the direction of the wave, and a *stretching* (outward acceleration) in the perpendicular transverse direction. These effects are simply those due to the *Weyl*-tensor (i.e. conformal) part of the Riemann curvature tensor. For an oscillatory wave, this squashing/stretching would be followed by the opposite effect (stretching/squashing), and so on repeatedly. To lowest order, we would expect that, overall, these squashing and stretching effects would ultimately cancel out when the wave has passed, but there would be a higher-order residual for which there would be a transverse *squashing* in *all* directions transverse to the direction to the wave. The latter would resemble the gravitational effect of the passage of a *matter* wave (e.g. electromagnetic), as given by a delta function in the *Ricci* tensor. These effects can be easily understood by considerations of ordinary classical lens optics, see [16], especially p.266. For a brief Planck-frequency gravitational wave, the resultant would appear to be an impulsive matter wav, carrying away the mass-energy of the erebon. (See [17] for this concept of an impulsive wave.) The effective space-time geometry is that of what could be referred to as an "impulsive Vaidya metric" (which is a Schwarzschild space-time converting itself to Minkowski space-time along an outgoing spherically symmetric null hypersurface, which would be an expanding spherical shell of delta-function Ricci tensor.

## 2. Erebon decays as an astrophysical probe.

If these effects of CCC are respected by nature, they should provide the potential for a powerful new way of exploring the universe. Conversely, if the stringent implications of this idea turn out *not* to be respected by nature, then it should not be too hard to establish this fact, and therefore refute the proposal that I am putting forward here. Of course, it might also be a possibility that something of the general nature of that which I am describing is indeed true,

but the model that I am describing needs modification from the basic picture being presented here. In any case, this proposal presents a possible resolution of the puzzle that appears to be posed by the correlations in the "noise" in the LIGO signals described at the beginning of this article.

Let us consider some relevant issues

(1) Since the erebon decays, are gravitational wave events, they should have an effect on a gravitational wave detector such as LIGO, though the signal is different from what is normally expected in the usual applications of Einstein theory, owing to the extremely high ($\approx 10^{43}$ Hz) frequency expected in the signal. Here the signal should be impulsive (instantaneous), with *inward* acceleration in all directions transverse to the direction of the wave.

(2) The signals would have a random character that would make them easily mistaken for irrelevant noise effects.

(3) If the theory is correct, these signals would provide a direct signal from dark matter concentrations, which could be compared with other estimates of where and how dense such distributions might be expected to be.

(4) Since it is supposed that all erebons have the same mass ($\approx 10^{-5}$g), where their decays provide point-like gravitational explosions of about the Planck energy ($\approx 2 \times 10^{9}$J $\approx 10^{19}$GeV) which is that of a sizable artillery shell, the strength of the received signal should provide a good measure of the distance of the source, if its radial velocity is otherwise known (or of its radial velocity if its distance is otherwise known).

(5) Any serious inconsistency with otherwise known facts would tell against the theory, so there is a considerable tightness in the overall scheme, making it extremely prone to observational disproof—which is a desirable feature of any serious scientific proposal.

(6) The above comments refer to what can be achieved by the use of data from the existing LIGO gravitational wave detectors. Of course more can be achieved when further detectors of the LIGO type are completed. Moreover, there is potentially a far greater scope for gravitational wave detectors based on BECs, as has been proposed by Ivette Fuentes and her collaborators [18-20], since, being very much smaller and far less expensive, could be produced in large numbers, giving a much better directional capacity than LIGO type detectors. Moreover, the BEC detectors would be more effective at high frequencies, and therefore better tuned to the impulsive waves here predicted from erebon decay.

The main uncertainty in the scheme as it stands appears to be the lifetime (or half-life) of erebon decay. We shall be seeing in Section 3 that, in CCC, there is a strong relation between this decay rate and the nature of the temperature fluctuations in our cosmic microwave background, which should, in principle give strong restrictions to what this decay can be, but this has not yet been looked at in any detail.

Let us now examine how erebon decays might influence the LIGO signals, appearing as "noise", but with the same time delay between the two detectors as have the black-hole encounter signals. The most obvious point is that the black hole pair would be expected to be in a galaxy, and that galaxy would, according to my proposal, be a source of erebon decay—where we bear in mind that the dark-matter density in a galaxy might be around a hundred times that in intergalactic space. Accordingly, that part of the "noise" which comes from the

erebon decay from that particular galaxy would indeed be correlated with the same time delay between the two detectors as would the signal from the black-hole encounter event.

Of course, in this description, the black-hole encounter is almost irrelevant, serving merely to locate an individual galaxy in the sky. To test this, on need merely settle on some prominent galaxy—say the Andromeda galaxy—and look for correlations in the noise that has a time delay between the two detectors that corresponds to sources in the direction of that particular galaxy. The black-hole encounter would be completely unnecessary for this, and one does not have to wait for such occurrences in order to test the proposal. Of course, the dark-matter distribution in our own galaxy would be a major contributor, in this proposed scheme, and thus should be a major contributor to the signals referred to here. Clearly the proposal that I am putting forward here makes many testable predictions, and it should not be hard to disprove it if it is wrong.

## 3. The ideas of CCC

According to CCC—the conformal cyclic cosmology proposal—the universe consists of a (perhaps infinite) succession of *aeons*, where each aeon originates with its own big bang and has an unending exponentially expanding future, consistent with a positive cosmological constant $\Lambda$, taken to have the same constant value from aeon to aeon, the Einstein $\Lambda$-equations holding throughout every aeon. The driving idea behind CCC is the second law of thermodynamics (abbreviated 2$^{nd}$ Law) in the particular form that it appears to have been presented to us by nature, namely that the initial extremely low entropy at the beginning of each aeon is manifested in the fact that while the matter and radiation are thermalized immediately following the big bang of each aeon, the gravitational degrees of freedom are taken to be completely suppressed at its big bang, so that the Weyl tensor $C_{abcd}$ vanishes there. In CCC, this "Weyl curvature hypothesis" is replaced by K.P.Tod's proposal [21] that each aeon's big bang can be smoothly conformally extended backwards into another space-time region. That region is, according to CCC, taken to be the remote future (conformal infinity) of a previous aeon.

There is a theorem, due to H.Friedrich [22] that tells us that, with positive $\Lambda$ and electromagnetism as the matter content, a smooth spacelike conformal future infinity will be a generic situation, so that the conformal continuation of the remote future of one aeon to some other space-time is a generic possibility, ad according to CCC, it would be to the big bang of a further aeon.

As this stands, this is merely a plausible geometrical picture, and it may be difficult to take it seriously as a physically acceptable model. However, when examined from the physical point of view, it does indeed hang together. The physics near the big bang is at an enormous temperature, so rest mass is completely dominated by kinetic motions, and so may be ignored.

Moreover our current picture of the distant future of the universe, according to Einstein's $\Lambda$-equations is a continuing exponential expansion, with matter density continually reducing and an entropy content completely dominated by supermassive black holes. Eventually (after some $10^{100}$ years), even these supermassive black holes will have all evaporated away by Hawking radiation, to leave an exceedingly rarified universe whose particle number content will be almost entirely in the form of photons, so the conditions of Friedrichs's theorem are

basically satisfied. There will, however, be a residue of a relatively small number of massive particles, and for the strict requirements of the theory one must apparently assume that there is a mechanism whereby in the hugely long-term future, mass eventually fades away, asymptotically—an admittedly unconventional hypothesis, but not contradicted by current observation, and not implausible on the basis of a particle physics which is extended from current ideas by the presence of a positive $\Lambda$.

Is it physically plausible that this enormously rarified and absurdly cold future infinity can be physically similar to the enormously dense and hot big bang of a succeeding aeon? Yes, indeed, if we may regard the relevant physics to be a conformally invariant one, where we bear in mind that not only are the equations of Maxwell conformally invariant, but so also are those of Yang-Mills, when mass can be ignored. This conformal freedom allows us to stretch out the hot big bang of the succeeding aeon and to squash down the remote future of the previous one—bearing in mind that energy and momentum scale in the exact inverse way to space and time. So hot becomes cold and dense becomes rarified upon conformal stretching.

A remark is necessary concerning the 2nd Law, as one might consider this problematic in a cyclic model. CCC's response is, first, to take note of the fact that, even now (according to the Bekenstein-Hawking entropy formula), by far the major contribution to the entropy in the universe is in supermassive black holes. This will vastly increase in the future. But ultimately, by Hawking radiation, all will disappear via Hawking evaporation. The issue here is that, in CCC, one must side with those who believe that information—or, more to the point, phase-space volume—must be destroyed by being obliterated at the hole-s singularity. As has often pointed out, this is in contradiction with quantum unitarity, but the view here is taken (in opposition to that of "many-worlds" proponents) that since unitarity is already violated in quantum measurement, then it is not a universal truth; moreover, the viewpoint will be taken here that in quantum measurement as well as with black-hole evaporation, quantum state reduction is a gravitational effect. This relates to the classicality of erebon particles, as already referred to in Section 1.

## 4. The equations of CCC

Of course one needs clear-cut equations to describe this transition from the remote future of one aeon to the big bang of the next. Such equations can be provided [6, Appendix A] [23], although there are still uncertainties, that will not concern us here, in certain detailed choices for ensuring uniqueness of evolution. Nevertheless, there are major issues concerning the evolution that are central to the current discussion.

The transition from each aeon to the next is described in terms of a "bandage metric" $g_{ab}$ that is defined throughout an open region $U$ containing the "crossover" hypersurface $X$ that connects the two aeons under consideration. Let $\hat{g}_{ab}$ be the physical metric for the aeon pre-crossover and $\breve{g}_{ab}$, the physical metric post-crossover. We have conformal factors $\Omega$ and $\omega$ so that

$$\hat{g}_{ab} = \Omega^2 g_{ab} \quad \text{and} \quad \breve{g}_{ab}, = \omega^2 g_{ab} ,$$

where the *reciprocal hypothesis*

$$\Omega = -\omega^{-1},$$

is taken to hold throughout $U$, and where $\omega$ is smooth, changing from negative to positive across $X$. Thus, $\Omega$ is infinite at $X$ and represents a "squashing down" of the remote future of the earlier aeon so as to fit smoothly on to the "stretched out" big bang of the subsequent aeon across $X$ within $U$ provided by $\omega$. Within the earlier aeon, the $\Omega$-field has no *physical* content, but we can take it as a kind of auxiliary field—a kind of "phantom field"—which allows us to express Einstein's equations in a conformally invariant framework, and when taken alongside the gravitational field allows us to extend the range of Einstein's theory so that it includes the crossover 3-surface $X$. There are no physical degrees of freedom in $\Omega$, beyond those already in the gravitational field. Yet it has an "energy tensor"

$$T_{ab}[\Omega] = \frac{1}{4\pi G}\Omega^3\{\nabla_a\nabla_b - \tfrac{1}{4}g_{ab}\nabla^c\nabla_c - \tfrac{1}{2}R_{ab} + \tfrac{1}{8}Rg_{ab}\}\Omega^{-1},$$

which is trace-free and divergence free and the Einstein $\Lambda$-equations become simply $T_{ab}=T_{ab}[\Omega]$, where $T_{ab}$ is the energy tensor of the matter (assumed to be massless).

However, on the future side of $X$, when the $\breve{g}_{ab}$ metric is used, the $\Omega$ behaves as an *actual* physical field (this being a consistent viewpoint because of the reciprocal hypothesis). This scalar field starts off as effectively massless, but would rapidly acquire a mass and would be expected to dominate over all other fields.

It is a hypothesis of CCC that this $\Omega$-field, when it acquires a mass, is in fact the *dark matter* field that provides the dominant matter content of the universe (where $\Lambda$ , or whatever else "dark energy" might be, does not count as matter). The gravitational degrees of freedom in the earlier aeon may be considered to be described by the Weyl tensor $\hat{C}_{abcd}$. We find that $\hat{C}_{abcd}$=0 at $X$ (which is in accordance with the Weyl curvature hypothesis), but gravitational wave signals in that earlier aeon do register at $X$ in the *normal derivative* of $\hat{C}_{abcd}$ across $X$. This information is not lost at crossover, but part of it (the "electric" part) is conveyed as higher derivatives in the $\Omega$-field (the "magnetic" part being encoded into a a 3-space conformal curvature—the Cotton tensor—for $X$).

Such signals would affect the CMB of the succeeding aeon in observable ways. Up until now, interest has been concentrated on the prospect of observing, in our current aeon, the effect of supermassive black-hole encounters in the aeon prior to ours. The remote future of the emitted gravitational waves should, according to CCC, register in the CMB of our aeon, as circular features, sometimes in concentric sets (indicating several such events occurring in the same galactic cluster in the previous aeon), and claims have been made that signals are indeed seen in the WMAP data [6], [7], and also in the Planck data [8], [9], though disputes remain as to the significance of these evident signals [10]. It may be remarked that these observations demonstrate that the sources of these signals are highly inhomogeneously distributed across the sky, which is very problematic for the conventional explanation in terms of quantum fluctuations in the inflaton field.

Of more relevance to the present discussion are the almost scale-invariant temperature fluctuations in the CMB. This scale invariance is one of the strongest motivations for a belief in an exponentially expanding inflationary phase in our very early universe. In CCC, there cannot be any such inflationary phase in our aeon, but the role of inflation is taken over, instead, by the exponentially expanding ultimate history of the previous aeon. The source of these fluctuations cannot be inflatons, since there is no inflation in CCC. However, their role is taken over by decaying particles pervading the whole of space in the previous aeon, and

these have to be the dark matter particles—the *erebons* referred to in earlier. Indeed these particles have ultimately to decay, in CCC, since we do not want them to build up from aeon to aeon. Their life-time has to be long, in comparison with the current age of our universe-aeon, so that the scale-invariance of the current exponential expansion, referred back to the aeon prior to ours can result in the observed close scale invariance in the CMB temperature fluctuations in our current aeon.

It is not the purpose of the current article to examine this feature in any detail. The concern, here, is to examine erebon decay in our own aeon, to see whether it might be an observable effect. The first point to take note of is the fact that no physical interactions other than gravity have been invoked in the above discussion. Erebons are scalar "conformal partners" to gravity. They would be expected to have no physical interactions other than gravitational. This is indeed consistent with their being the particles that constitute dark matter, since dark matter does actually appear to interact only gravitationally with other forms of matter. Since only gravity is involved, this suggests that the erebons should have a *Planck mass*, or a simple multiple of that mass, as put forward in Section 1. Finally, the fact that they should behave as classical particles (as proposed in Section 1), and, more importantly, as their decay should be in the form as a classical extremely high-frequency classical gravitational wave is crucial for the picture. A Planck-energy *quantum* particle would have a catastrophic effect on a detector which measured it, since the entire artillery-shell energy would be transferred to the detector upon measurement!

My thanks go to Iwo Bialynicki-Birula, Paul Tod, and most particularly to Krzysztof Meissner, discussions with whom led to the idea that dark-matter decay in the previous aeon might be responsible for the near scale invariant CMB temperature fluctuations, and to Ivette Fuentes, with whom discussions of possible detection of such decay in our own aeon using her proposed BEC detectors led me to the unusual line of thinking suggested here. I am also grateful for financial assistance through a personal endowment from J.P. Moussouris.

## References

[1] Cresswell, J., vonHausegger, S., Jackson, A.D., Hao Liu, and Naselsky, P. (2017) On the time lags of the LIGO signals. arXiv:1706.04191v1 [astro-ph.IM] 13 Jun 2017

[2] See e.g. Harry I, http://www.preposterousuniverse.com/blog/2017/06/18/a-response-to-on-the-time-lags-of-the-ligo-signals-guest-post/; Cresswell et al., http://www.nbi.ku.dk/gravitational-waves/gravitational-waves.html.

[3] Penrose, R., (2017) Dark matter decay? Possible observational tests—according to conformal cyclic cosmology. ESI Vienna, 20 June 2017 http://quark.itp.tuwien.ac.at/~grumil/ESI2017/talks/penrose.mp4

[4] Penrose, R., (2017) Correlated "noise" in LIGO Gravitational wave signals (submitted to Physical Review Letters)

[5] Penrose, R. (2010) *Cycles of Time: An Extraordinary New View of the Universe.* (Bodley Head, London). (ISBN 978-0-224-08036-1; 9780099505945); Penrose, R. (2014) On the gravitization of quantum mechanics 2: conformal cyclic cosmology. *Found Phys* **44**:873–890.

[6] Gurzadyan, V.G. and Penrose R. (2013) On CCC-predicted concentric low-variance circles in the CMB sky. *Eur. Phys. J. Plus* (2013) 128: 22 DOI 10.1140/epjp/i2013 13022-4.

[7] Meissner, K. A., Nurowski, P. and Ruszczycki, B. (2013) Structures in the microwave background radiation, *Proc. R. Soc.* **A469**:2155, 20130116, arXiv:1207.2498[astro-ph.CO]
[8] Gurzadyan,V.G. and Penrose, R. (2016) CCC and the Fermi paradox *Eur. Phys. J. Plus* (2016) 131: 11.
[9] An, D., Meissner, K.A., Nurowski, P. (2013) Structures in the Planck map of the CMB arXiv:1307.5737 [astro-ph.CO]. (Submitted on 22 Jul 2013.)
[10] DeAbreu, A., Contreras, C., and Scott, D. (2015) JCAP, **12**, 031
[11] Károlyházy, F. (1966) Gravitation and quantum mechanics of macroscopic bodies, *Nuovo Cim.* **A42**, 390. Károlyházy, F. (1974) Gravitation and quantum mechanics of macroscopic bodies, *Magyar Fizikai Polyoirat* **12**, 24.
[12] Diósi, L. (1984) Gravitation and quantum mechanical localization of macro-objects Phys. Lett. 105A, 199-202. Diósi, L. (1989) Models for universal reduction of macroscopic quantum fluctuations, Phys. Rev. A40, 1165-74.
[13] Penrose, R. (1996) On gravity's role in quantum state reduction *Gen. Rel. Grav.* **28**, 581-600. Penrose, R. (1993) Gravity and quantum mechanics. In *General Relativity and Gravitation 13. Part 1: Plenary Lectures 1992. Proceedings of the Thirteenth International Conference on General Relativity and Gravitation held at Cordoba, Argentina, 28 June - 4 July 1992.* Eds. R.J.Gleiser, C.N.Kozameh, and O.M.Moreschi (Inst. of Phys. Publ. Bristol & Philadelphia) pp.179-89.
[14] Penrose, R. (2014) On the gravitization of quantum mechanics 1: quantum state reduction. *Found Phys* **44**:557–575.
[15] Percival, I.C. (1995) Quantum spacetime fluctuations and primary state diffusion. *Proc. R. Soc. Lond.* **A451**, 503-13.
[16] Penrose, R. (1966) General-relativistic energy flux and elementary optics, in *Perspectives in Geometry and Relativity*, ed. B. Hoffmann (Indiana University Press, Bloomington) pp.259-274.
[17] Penrose, R. (1972) The Goemetry of Impulsive Gravitational Waves, in *General Relativity, Papers in Honour of J.L. Synge.* ed. L.O'Raifeartaigh (Clarendon Press, Oxford), pp. 101-15.
[18] Sabín, C., Bruschi, D.E., Ahmadi, M. and Fuentes, I. (2014) *New Journal of Physics* **16**, 085003.
[19] Sabín, C., Kohlrus, J., Bruschi, D. E., and Fuentes, I. (2016) *EPJ Quantum Technology* **3**, 8.
[20] Howl, R., Sabín, C., Hackermüler, L., Fuentes, I (2016) arXiv:1612.01931.
[21] Tod, K.P. (2003) Isotropic cosmological singularities: other matter models. *Class. Quantum Grav.* **20** 521-534 doi:10.1088/0264-9381/20/3/309
[22] Friedrich, H. (1998) Einstein's equation and conformal structure. In *The Geometric Universe; Science, Geometry, and the Work of Roger Penrose*, eds. S.A. Huggett, L.J. Mason, K.P. Tod, S.T. Tsou, and N.M.J. Woodhouse (Oxford Univ. Press, Oxford).
[23] Szabados, L.B. and Tod, K.P. (2015) *Classical and Quantum Gravity* **32**, 205011.

arXiv:quant-ph/0210001v1 30 Sep 2002

# Towards quantum superpositions of a mirror

William Marshall[1,2], Christoph Simon[1], Roger Penrose[3,4] and Dik Bouwmeester[1,2]
[1] *Department of Physics, University of Oxford, Oxford OX1 3PU, United Kingdom*
[2] *Department of Physics, University of California, Santa Barbara, CA 93106*
[3] *Center for Gravitational Physics and Geometry, The Pennsylvania State University, University Park, PA 16802*
[4] *Department of Mathematics, University of Oxford, Oxford OX1 3LB, United Kingdom*
(Dated: February 1, 2008)

We propose a scheme for creating quantum superposition states involving of order $10^{14}$ atoms via the interaction of a single photon with a tiny mirror. This mirror, mounted on a high-quality mechanical oscillator, is part of a high-finesse optical cavity which forms one arm of a Michelson interferometer. By observing the interference of the photon only, one can study the creation and decoherence of superpositions involving the mirror. All experimental requirements appear to be within reach of current technology.

In 1935 Schrödinger pointed out that according to quantum mechanics even macroscopic systems can be in superposition states [1]. The quantum interference effects are expected to be hard to detect due to environment induced decoherence [2]. Nevertheless there have been proposals on how to create and observe macroscopic superpositions in various systems [3–5], and experiments demonstrating superposition states of superconducting devices [6] and fullerene molecules [7]. One long-term motivation for this kind of experiment is the question of whether unconventional decoherence processes such as gravitationally induced decoherence or spontaneous wave-function collapse [8, 9] occur.

Here we present a scheme that is close in spirit to Schrödinger's original discussion. A small quantum system (a photon) is coupled to a large system (a mirror) in order to create a macroscopic superposition. The basic principle of the experiment as described in Ref. [9] grew out of discussions in 1997 [10]. It consists of a Michelson interferometer in which one arm has a tiny moveable mirror. The radiation pressure of a single photon displaces the tiny mirror. The initial superposition of the photon being in either arm causes the system to evolve into a superposition of states corresponding to two distinct locations of the mirror. In the present proposal a high-finesse cavity is used to enhance the interaction between the photon and the mirror. The observed interference of the photon allows one to study the creation of coherent superposition states periodic with the motion of the mirror. We perform a detailed analysis of the requirements for such an experiment.

The proposed setup is shown in Fig. 1. In the cavity in arm A one of the mirrors is very small and attached to a micromechanical oscillator, similar to the cantilevers in atomic force microscopes. While the photon is in the cavity, it exerts radiation pressure on the mirror. Under the conditions that the oscillator period is much longer than the photon roundtrip time, and the amplitude of the mirror's motion is small compared to the cavity length,

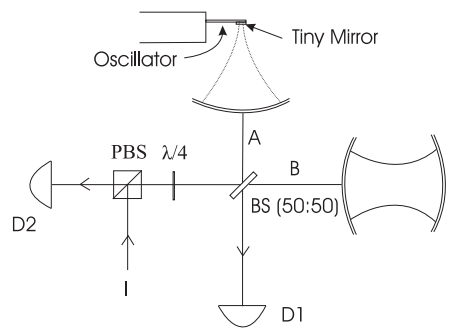



FIG. 1: The proposed setup: a Michelson interferometer for a single photon, where in each arm there is a high-finesse cavity. The cavity in arm A has a very small end mirror mounted on a micro-mechanical oscillator. The single photon comes in through I. If the photon is in arm A, the motion of the small mirror is affected by its radiation pressure. The photon later leaks out of either cavity and is detected at D1 or D2.

the system can be described by a Hamiltonian [11]

$$H = \hbar\omega_c a^\dagger a + \hbar\omega_m b^\dagger b - \hbar g a^\dagger a(b + b^\dagger), \qquad (1)$$

where $\omega_c$ is the frequency of the photon in the cavity for the (empty) cavity length $L$, $a$ is the creation operator for the photon, $\omega_m$ and $b$ are the frequency and phonon creation operator for the center of mass motion of the mirror, and the coupling constant is $g = \frac{\omega_c}{L}\sqrt{\frac{\hbar}{2M\omega_m}}$, where $M$ is the mass of the mirror. To start with, let us suppose that initially the photon is in a superposition of being in either arm $A$ or $B$, and the mirror is in a coherent state $|\beta\rangle = e^{-|\beta|^2/2}\sum_{n=0}^{\infty}\frac{\beta^n}{\sqrt{n!}}|n\rangle$, where $|n\rangle$ are the eigenstates of the harmonic oscillator. Then the initial state is $|\psi(0)\rangle = \frac{1}{\sqrt{2}}(|0\rangle_A|1\rangle_B + |1\rangle_A|0\rangle_B)|\beta\rangle$. After a time $t$ the state of the system will be given by [12]

$$|\psi(t)\rangle = \frac{1}{\sqrt{2}}e^{-i\omega_c t}(|0\rangle_A|1\rangle_B|\beta e^{-i\omega_m t}\rangle + e^{i\kappa^2(\omega_m t - \sin\omega_m t)}|1\rangle_A|0\rangle_B|\beta e^{-i\omega_m t} + \kappa(1 - e^{-i\omega_m t})\rangle), \qquad (2)$$

where $\kappa = g/\omega_m$. In the second term on the r.h.s. the motion of the mirror is altered by the radiation pressure of the photon in cavity $A$. The parameter $\kappa$ quantifies the displacement of the mirror in units of the size of the coherent state wavepacket. In the presence of the photon the mirror oscillates around a new equilibrium position determined by the driving force.

The maximum possible interference visibility for the photon is given by twice the modulus of the off-diagonal element of the photon's reduced density matrix. By tracing over the mirror one finds from Eq. (2) that the off-diagonal element has the form

$$\frac{1}{2}e^{-\kappa^2(1-\cos\omega_m t)}e^{i\kappa^2(\omega_m t-\sin\omega_m t)+i\kappa\mathrm{Im}[\beta(1-e^{i\omega_m t})]} \quad (3)$$

where Im denotes the imaginary part. The first factor is the modulus, reaching a minimum after half a period at $t = \pi/\omega_m$, when the mirror is at its maximum displacement. The second factor gives the phase, which is identical to that obtained classically due to the varying length of the cavity. For general $t$ the phase depends on $\beta$, i.e. the initial conditions of the mirror. However, the effect of $\beta$ averages out after every full period.

In the absence of decoherence, after a full period, the system is in the state $\frac{1}{\sqrt{2}}(|0\rangle_A|1\rangle_B + e^{i\kappa^2 2\pi}|1\rangle_A|0\rangle_B)|\beta\rangle$, such that the mirror is again disentangled from the photon. Full interference can be observed if the photon is detected at this time. If the environment of the mirror "remembers" that the mirror has moved, then, even after a full period, the photon will still be entangled with the mirror's environment, and thus the interference for the photon will be reduced. Therefore the setup can be used to measure the decoherence of the mirror.

In practice the mirror will be in a thermal state, which can be written as a mixture of coherent states $|\beta\rangle$ with a Gaussian probability distribution $(1/\pi\bar{n})e^{-|\beta|^2/\bar{n}}$, where $\bar{n} = 1/(e^{\hbar\omega_m/kT}-1)$ is the mean thermal number of excitations. In order to determine the expected interference visibility of the photon at a time $t$ for an initial mirror state which is thermal, one has to average the off-diagonal element Eq. (3) over $\beta$ with this distribution. The result is

$$\frac{1}{2}e^{-\kappa^2(2\bar{n}+1)(1-\cos\omega_m t)}e^{i\kappa^2(\omega_m t-\sin\omega_m t)}. \quad (4)$$

As a consequence of the averaging of the $\beta$-dependent phase in Eq. (3), the off-diagonal element now decays on a timescale $1/(\kappa\omega_m\sqrt{\bar{n}})$ after $t = 0$, i.e. very fast for the experimentally relevant case of $\kappa \sim 1$ and large $\bar{n}$. However, remarkably, it still exhibits a revival [4] at $t = 2\pi/\omega_m$, when photon and mirror become disentangled and the phase in Eq. (3) is independent of $\beta$, such that the phase averaging does not reduce the visibility. This behaviour is shown in Fig. 2. The magnitude of the revival is reduced by any decoherence of the mirror.

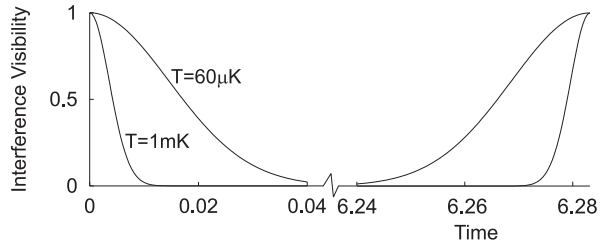



FIG. 2: Time evolution of the interference visibility of the photon over one period of the mirror's motion for $\kappa = 1$, $T = 1$mK and $T = 60\mu$K. The time is given in units of $1/\omega_m$. The visibility decays very fast after $t = 0$, but in the absence of decoherence there is a sharp revival of the visibility after a full period. The width of each peak scales like $1/\sqrt{T}$.

The revival demonstrates the coherence of the superposition state that exists at intermediate times. For $\kappa^2 \gtrsim 1$ the state of the system is a superposition involving two distinct mirror positions. More precisely, for a thermal mirror, the state of the system is a mixture of such superpositions. However, this does not affect the fundamentally non-classical character of the state. We now discuss the experimental requirements for achieving such a superposition and observing its recoherence at $t = 2\pi/\omega_m$.

Firstly, we require $\kappa^2 \gtrsim 1$, which physically means that the momentum kick imparted by the photon has to be larger than the initial quantum uncertainty of the mirror's momentum. Let $N$ denote the number of roundtrips of the photon in the cavity during one period of the mirror's motion, such that $2NL/c = 2\pi/\omega_m$. This allows us to rewrite the condition $\kappa^2 \gtrsim 1$ as

$$\frac{2\hbar N^3 L}{\pi c M\lambda^2} \gtrsim 1, \quad (5)$$

where $\lambda$ is the wavelength of the light. The factors entering Eq. (5) are not all independent. The achievable $N$, which is determined by the quality of the mirrors, and the minimum mirror size (and hence $M$) both depend strongly on $\lambda$. The mirror's lateral dimension should be an order of magnitude larger than $\lambda$ to limit diffraction losses. Its thickness in order to achieve sufficiently high reflectivity depends on $\lambda$ as well.

Eq. (5) allows one to compare the viability of different wavelength ranges. While the highest values for $N$ are achievable for microwaves (up to $10^{10}$), this is counteracted by their cm wavelengths. On the other hand there are no good mirrors for highly energetic photons. The optical regime seems optimal. Here we propose an experiment with $\lambda$ around 630 nm.

The cavity mode needs to have a very narrow focus on the tiny mirror, which requires the other cavity end mirror to be large due to beam divergence. The maximum cavity length is therefore limited by the difficulty of making large high quality mirrors. We propose a cavity length of 5 cm, and a small mirror size of $10 \times 10 \times 10$

microns, leading to a mass of order $5 \times 10^{-12}$ kg.

Such a small mirror on a mechanical oscillator can be fabricated by coating a silicon cantilever with alternating layers of $SiO_2$ and a metal oxide. The best current mirrors are made in this way. A larger silicon oscillator has been coated with $SiO_2/Ta_2O_5$ and used as part of a high-finesse cavity in Ref. [13].

For the above dimensions the condition (5) is satisfied for $N = 5.6 \times 10^6$. Correspondingly, we are aiming for a photon loss per reflection not larger than $3 \times 10^{-7}$, about a factor of 4 below the best reported values for such mirrors [14], and for a transmission of $10^{-7}$, consistent with the $10\mu$m mirror thickness [15]. For these values, about 1% of the photons are still left in the cavity after a full period of the mirror. For the above values of $N$ and $L$ one obtains a frequency $\omega_m = 2\pi \times 500$ Hz. This leads to a quantum uncertainty for the mirror position of order $10^{-13}$ m, which corresponds to the necessary displacement in the superposition in order to achieve $\kappa^2 \sim 1$.

The fact that a relatively large $L$ is needed to satisfy Eq. (5) implies that the creation of superpositions following the related proposal of Ref. [4], which relies on a micro-cavity, imposes requirements that are probably beyond the reach of current technology. A large $L$ is helpful because, for a given $N$, it allows one to use a lower frequency $\omega_m$, and thus a more weakly bound mirror that is easier to displace by the photon.

Secondly, the requirement of observing the revival puts a bound on the acceptable environmental decoherence. To estimate the expected decoherence we model the mirror's environment by an (Ohmic) bath of harmonic oscillators. The effect of this environment can approximately be described by a decoherence rate $\gamma_D = \gamma_m kTM(\Delta x)^2/\hbar^2$ governing the decay of off-diagonal elements between different mirror positions [2]. Here $\gamma_m$ is the damping rate for the mechanical oscillator, $T$ is the temperature of the environment, which is constituted mainly by the internal degrees of freedom of the mirror and cantilever, and $\Delta x$ is the separation of two coherent states that are originally in a superposition. This approximation is strictly valid only for times much longer than $2\pi/\omega_m$ and for $\Delta x$ large compared to the width of the individual wavepackets. A more rigorous description would follow Ref. [16] and take into account the movement of the mirror. Our analysis indicates that the order of magnitude of the decoherence is well captured by $\gamma_D$. Assuming that the experiment achieves $\kappa^2 \gtrsim 1$, i.e. a separation by the size of a coherent state wavepacket, $\Delta x \sim \sqrt{\frac{\hbar}{M\omega_m}}$, the condition $\gamma_D \lesssim \omega_m$ can be cast in the form

$$Q \gtrsim \frac{kT}{\hbar\omega_m} = \bar{n}, \tag{6}$$

where $Q = \omega_m/\gamma_m$ is the quality factor of the mechanical oscillator. Bearing in mind that $Q \gtrsim 10^5$ has been achieved [17] for silicon cantilevers of approximately the right dimensions and frequency, this implies that the temperature has to be 3 mK or less.

Thirdly, the stability requirements for the experiment are very strict. The phase of the interferometer has to be stable over the whole measurement time. This means that the distance between the large cavity end mirror and the equilibrium position [18] of the small mirror has to be stable to of order $\lambda/20N = 0.6 \times 10^{-14}$m. The required measurement time can be determined in the following way. A single run of the experiment starts by sending a weak pulse into the interferometer, such that on average 0.1 photons go into either cavity. This probabilistically prepares a single-photon state as required to a good approximation. The two-photon contribution has to be kept low because it causes noise in the interferometer. From Eq. (4) the width of the revival peak is $2/\kappa\omega_m\sqrt{\bar{n}}$. This implies that only a fraction $\sim 1/\pi\sqrt{\bar{n}}$ of the remaining photons will leak out in the time interval of the revival. It is therefore important to work at the lowest possible temperature. A temperature of 60 $\mu$K has been achieved with a nuclear demagnetization cryostat [19].

Considering the required low value of $\omega_m$ and the fact that approximately 1% of the photons remain after a full period for the assumed loss, this implies a detection rate of approximately 100 photons per hour in the revival interval, given a detection efficiency of 70 %. Thus a measurement time of order 10 minutes should give statistically significant results. After one such measurement period the interferometer can be readjusted, and the experiment can be repeated. Stability of order $10^{-13}$m/min for an STM inside at 8 K was achieved with a rather simple suspension [20]. Gravitational wave observatories using interferometers also require very high stability in order to have an length sensitivity of $10^{-19}$m over timescales of a ms or greater [21], for arm lengths of order 1 km.

The experiment requires ultra-high vacuum conditions in order to ensure that events where an atom hits the cantilever are sufficiently rare not to cause significant errors, which is at the level of about 5/s. Background gas particle densities of order $100/\text{cm}^3$ have been achieved [22] and are sufficient for our purposes.

After every single run of the experiment the mirror has to be damped to reset it to its initial (thermal) state. This could be done by electric or magnetic fields, e.g. following Ref. [23], where a Nickel sphere was attached to the cantilever, whose $Q$ could then be changed by three orders of magnitude by applying a magnetic field.

There are several ways in which the total measurement time could be decreased, thus relaxing the stability requirements. Switchable cavity mirrors would allow coupling the photon into and out of the cavity at any desired time, instead of relying on postselection. This allows the data collection in the revival interval to be increased by three orders of magnitude. The transmission of a distributed Bragg reflector (DBR) can be changed

by altering the refractive index of one of the two materials. Highly reflective mirrors have been fabricated from alternating layers of GaAs and AlAs [24]. The refractive index of GaAs can be changed dramatically by optical pumping [25]. A change from 3.6 to 3.15 on a ps timescale was observed at a probe wavelength of 830nm for pump intensities in the range 1 kJ/m$^2$ (just below the damage threshold of the material). Such a large change in refractive index would allow efficient optical switching [26] and is a promising avenue that deserves experimental investigation. The main open question is whether the absorption of the materials [27] can be made low enough to match our very strict demands.

Since the width of the revival peak scales like $1/\sqrt{T}$, the required measurement time can also be decreased by decreasing the temperature below 60 $\mu$K. Passive cooling techniques may be improved. In addition, active cooling of mirror oscillators has been proposed [28], and even implemented experimentally for a large mirror [29]. The mirror's movement is sensed via the output light of an optical cavity, and a feedback loop is used to damp the motion. Applying the theory of Ref. [28] to our mirror, one finds that cooling even to the ground state of the center of mass motion is conceivable. This would reduce the required measurement time, and thus the stability requirements, by a factor of 50.

In principle the proposed setup has the potential to test wave function reduction models, in particular the one of Ref. [9]. Based on [9] and [16], we estimate that an improvement of the ratio $Q/T$ by about five orders of magnitude from the values discussed in this paper ($Q = 10^5$ and $T = 60\mu$K) would make the predicted wavefunction decoherence rate comparable to the environmental decoherence rate. Improvements in $Q$ are certainly conceivable. In particular, $Q$ is known to increase with decreasing temperature [17]. Active cooling methods for the mirror's center of mass motion as described above could in principle also be used to indirectly cool the mirror's internal degrees of freedom. Whether such sympathetic cooling is a realistic avenue is in itself a fascinating subject for further study.

We have performed a detailed study of the experimental requirements for the creation and observation of quantum superposition states of a mirror consisting of $10^{14}$ atoms, approximately nine orders of magnitude more massive than any superposition observed to date. Our analysis suggests that, while very demanding, this goal appears to be in reach of current technology. It is remarkable that a tabletop experiment has the potential to test quantum mechanics in an entirely new regime. Preliminary experiments on components of the proposal are currently under way.

This work was supported by the E.U. (IST$-1999-10033$). W.M. is supported by EPSRC (award no. 00309297). C.S. is supported by a Marie Curie Fellowship of the E.U. (no. HPMF-CT-2001-01205). R.P. thanks the NSF for support under contract 00-90091 and the Leverhulme Foundation for an Emeritus Fellowship. We would like to thank S. Bose, M. Davies, R. Epstein, T. Knuuttila, R. Lalezari, J. Pethica, and J. Roberts for useful discussions.

[1] E. Schrödinger, Die Naturwissenschaften **23**, 807 (1935).
[2] D. Giulini *et al.*, *Decoherence and the Appearance of a Classical World in Quantum Theory* (Springer, Berlin, 1996); W.H. Zurek, Phys. Today **44**, 36 (1991).
[3] J. Ruostekoski, M. J. Collett, R. Graham, and D. F. Walls, Phys. Rev. A **57**, 511 (1998); J. I. Cirac, M. Lewenstein, K. Molmer, and P. Zoller, Phys. Rev. A **57**, 1208 (1998).
[4] S. Bose, K. Jacobs, and P.L. Knight, Phys. Rev. A **59**, 3204 (1999).
[5] A.D. Armour, M.P. Blencowe, and K.C. Schwab, Phys. Rev. Lett. **88**, 148301 (2002).
[6] C.H. van der Wal *et al.*, Science **290**, 773 (2000); J.R. Friedman, V. Patel, W. Chen, S.K. Tolpygo, and J.E. Lukens, Nature **406**, 43 (2000).
[7] M. Arndt *et al.*, Nature **401**, 680 (1999).
[8] G.C. Ghirardi, A. Rimini and T. Weber, Phys. Rev. D **34**, 470 (1986); G.C. Ghirardi, P. Pearle, and A. Rimini, Phys. Rev. A **42**, 78 (1990); I.C. Percival, Proc. Roy. Soc. London, Ser. A **447**, 189 (1994); D.I. Fivel, Phys. Rev. A **56**, 146 (1997); L. Diosi, Phys. Rev. A **40**, 1165 (1989).
[9] R. Penrose, in A. Fokas *et al.* (Eds.), *Mathematical Physics 2000* (Imperial College, London, 2000).
[10] J. Schmiedmayer, R. Penrose, D. Bouwmeester, J. Dapprich, H. Weinfurter, A. Zeilinger, private communication (1997). See also D. Bouwmeester *et al.*, in N. Dadhich and J. Narlikar (Eds.), *Gravitation and Relativity: At the turn of the Millennium* (IUCAA, Pune, 1998).
[11] C.K. Law, Phys. Rev. A **51**, 2537 (1994); C.K. Law, Phys. Rev. A **49**, 433 (1993).
[12] S. Mancini, V.I. Man'ko, and P. Tombesi, Phys. Rev. A **55**, 3042 (1997); S. Bose, K. Jacobs, and P.L. Knight, Phys. Rev. A **56**, 4175 (1997).
[13] I. Tittonen *et al.*, Phys. Rev. A **59**, 1038 (1999).
[14] G. Rempe, R.J. Thompson, H.J. Kimble, and R. Lalezari, Opt. Lett. **17**, 363 (1992).
[15] C.J. Hood, H.J. Kimble, and J. Ye, Phys. Rev. A **64**, 033804 (2001).
[16] W.T. Strunz and F. Haake, quant-ph/0205108 (2002)
[17] H.J. Mamin and D. Rugar, Appl. Phys. Lett. **79**, 3358 (2001).
[18] As we have seen above, the thermal fluctuations of the small mirror around this equilibrium position (which are of order $10^{-12}$ m for 60 $\mu$K) lead to a narrowing of the visibility peaks, but do not destroy the revival.
[19] W. Yao *et al.*, J. Low Temp. Phys. **120**, 121 (2000).
[20] B.C. Stipe, M.A. Rezaei and W. Ho., Rev. Sci. Inst. **70**, 137 (1999).
[21] S. Rowan and J. Hough, Living Review in Relativity, **3**, 2000-3 (2000),
[22] G. Gabrielse *et al.*, Phys. Rev. Lett. **65**, 1317 (1990).
[23] K. Wago, D. Botkin, C.S. Yannoni, and D. Rugar, Appl. Phys. Lett. **72**, 2757 (1998).
[24] T.E. Sale, *Vertical Cavity Surface Emitting Lasers* (Re-

search Studies Press, Baldock, 1995).
[25] L. Huang, J.P. Callan, E.N. Glezer, and E. Mazur, Phys. Rev. Lett. **80**, 185 (1998).
[26] M. Scalora, J.P. Dowling, C.M. Bowden, and M.J. Bloemer, Phys. Rev. Lett. **73**, 1368 (1993).
[27] M.D. Sturge, Phys. Rev. **127**, 768 (1962).
[28] S. Mancini, D. Vitali, and P. Tombesi, Phys. Rev. Lett. **80**, 688 (1998).
[29] P.F. Cohadon, A. Heidmann, and M. Pinard, Phys. Rev. Lett. **83**, 3174 (1999).

# Beyond the Doubting of a Shadow

## A Reply to Commentaries on *Shadows of the Mind*

**Roger Penrose**
Mathematical Institute
24-29 St. Giles
Oxford OX1 3LB
U.K.

Copyright (c) Roger Penrose 1996

PSYCHE, 2(23), January 1996
http://psyche.cs.monash.edu.au/v2/psyche-2-23-penrose.html

KEYWORDS: artificial intelligence, free will, Gödel's theorem, mathematics, microtubules, Platonism, quantum mechanics.

REPLIES TO:

1. Bernard J. Baars: Can physics provide a theory of consciousness?
2. David J. Chalmers: Minds, machines, and mathematics
3. Solomon Feferman: Penrose's Gödelian argument
4. Stanley A. Klein: Is quantum mechanics relevant to understanding consciousness?
5. Tim Maudlin: Between the motion and the act....
6. John McCarthy: Awareness and understanding in computer programs
7. Daryl McCullough: Can humans escape Gödel?
8. Drew McDermott: [STAR] Penrose is wrong
9. Hans Moravec: Roger Penrose's gravitonic brains

CONTENTS

1. General remarks
2. Some technical slips in *Shadows*
3. The central new argument of *Shadows*
4. The "bare" Gödelian case
5. Gödel's "theorem-proving machine"
6. The issue of errors
7. The "unknowability" issue
8. AI and MJC
9. Mathematical Platonism
10. What has Gödel's theorem to do with physics?
11. How could physics actually help?
12. State-vector reduction
13. Free will

14. Some remarks on biology
15. What is consciousness?

# 1. General Remarks

1.1 I am glad to have this opportunity to address some of the criticisms that have been aimed at arguments in my book *Shadows of the Mind* (henceforth *Shadows*). I hope that in the following remarks I am able to remove some of the confusions and misunderstandings that still surround the arguments that I tried to make in that book - and also that we may be able to move forward from there.

1.2 In the accompanying PSYCHE articles, the great majority of the commentators' specific criticisms have been concerned with the purely logical arguments given in Part 1 of *Shadows*, with comparatively little reference being made to the physical arguments given in Part 2 - and virtually none at all to the biological ones.<1> This is not unreasonable if it is regarded that the entire rationale for my physical and biological arguments stands or falls with my purely logical arguments. Although I do not entirely agree with this position - since I believe that there are strong motivations from other directions for the kinds of physical and biological action that I have been promoting in *Shadows* - I am prepared to go along with it for the moment. Thus, most of my remarks here will be concerned with the implications of Gödel's theorem, and with the claims made by many of my critics that my arguments do not actually establish that there must be a noncomputational ingredient in human conscious thinking.

1.3 In replying to these arguments, I should first point out that, very surprisingly, almost none of the commentators actually addresses what I had regarded as the central (new) core argument against the computational modelling of mathematical understanding! Only Chalmers actually draws attention to it, and comments in detail on this argument, remarking that "most commentators seem to have missed it".<2> Chalmers also remarks that "it is unfortunate that this argument was so deeply buried". I apologize if this appears to have been the case; but I am also very puzzled, since its essentials are summarized in the final arguments of "Reductio ad absurdum - a fantasy dialogue", which is the section of *Shadows* (namely Section 3.23) that readers are particularly directed towards. This section is referred to also by McDermott and by Moravec, but neither of these commentators actually addresses this central argument explicitly, and nor do any of the other commentators. This is particularly surprising in the case of McCullough, as he is concerned with some of the subtleties of the logic involved, and also of Feferman, in view of his very carefully considered logical discussion.

1.4 It would appear, therefore, that I have an easy solution to the problem of replying to all nine commentators. All I need do is show why the ingenious argument put forward by Chalmers (based partly on McCullough's very general considerations) as a counter to my central argument is in fact (subtly) invalid! However, I am sure that this mode of procedure would satisfy none of the other commentators, and many of them also have interesting other points to make which need commenting upon. Accordingly, in the

following remarks, I shall attempt to address all the serious points that they do bring up. My reply to this main argument of Chalmers (partly dependent upon that of McCullough) will be given in Section 3, but it will be helpful first to precede this by addressing, in Section 2, the significant logical points that are raised by Feferman in his careful commentary.

# 2. Some Technical Slips in *Shadows*

2.1 Feferman quite correctly draws attention to some inaccuracies in *Shadows* with regard to certain logical technicalities. The most significant of these (in fact, the only really significant one for my actual arguments) concerns a misunderstanding on my part with regard to the assertion of omega-consistency of a formal system F, which I had chosen to denote by the symbols Omega(F), and its relation to Gödel's first incompleteness theorem. (As it happens, two others before Feferman had also pointed out this particular error to me.) As Feferman says, the assertion that some particular formal system is "omega-consistent" is certainly not of the form of a PI_1-sentence (i.e. not of the form of an assertion: "such-and-such a Turing computation never halts" - I call these "P-sentences" from here on). This much I should have been (and essentially was) aware of, despite the fact that in the first two printings of *Shadows*, p.96 I made the assertion that Omega(F) is a P-sentence. The fact of the matter was that I had somehow (erroneously) picked up the belief that the statement that Gödel originally exhibited in his famous first incompleteness theorem was *equivalent* to the omega-consistency of the formal system in question, not that it merely *followed* from this omega-consistency. Accordingly, I had imagined that for some technical reason I did not know of, this omega-consistency must actually be equivalent (for sufficiently extensive systems F) to the particular assertion "C_k(k)" that I had exhibited in Section 2.5, when the rules of the formal system F are translated into the algorithm A. Accordingly, I had mistakenly believed that Omega(F) must, for some subtle reason (unknown to me), be equivalent to the P-sentence C_k(k) (at least for sufficiently extensive systems F).

2.2 This error affects none of the essential arguments of the book but it is unfortunate that in various parts of Chapter 3, and most particularly in the "fantasy dialogue" in Section 3.23, the notation "Omega(F)" is used in circumstances where I had intended this to stand for the actual P-sentence C_k(k). In later printings of *Shadows*, this error has been corrected: I use the Gödel sentence G(F) (which asserts the consistency of F and is a P-sentence) in place of Omega(F). It is in any case much more appropriate to use G(F) in the arguments of Chapter 3, rather than Omega(F), and I agree with Feferman that the introduction of "Omega(F)" was essentially a red herring. In fact, the presentation in *Shadows* would have usefully simplified if omega-consistency had not even been mentioned.

2.3 The next most significant point of inaccuracy - or rather imprecision - in *Shadows* that Feferman brings up is that there is a discrepancy between different notions of the term "sound" that I allude to in different parts of the book. (This is actually quite an important issue, in relation to some of the discussion to follow, and I shall need to return

to it later in Section 3.) His point is, essentially, that in some places I need make use of the soundness of a formal system only in the limited sense of its capacity to assert the truth of certain P-sentences, whereas in other places I am actually referring to soundness in a more comprehensive sense, where it applies to other types of assertion as well. I agree that I should have been more careful about such distinctions. In fact, it is the weaker notion of soundness that would be sufficient for all the "Gödelian" arguments that I actually use in Part 1 of *Shadows*, though for some of the more philosophical discussions, I had in mind soundness in a stronger sense. (This stronger sense is not needed on pp. 90-92 if omega-consistency is dropped; nor is it needed on p.112, the weaker notion of soundness now being equivalent to consistency.)

2.4 Basically, I am happy to agree with all the technical criticisms and corrections that Feferman refers to in his section discussing my treatment of the logical facts". (I should attempt a point of clarification concerning his puzzlement as to why I should make the "strange" and "trivial" assertions he refers to on p.112. No doubt I expressed myself badly. The point that I was attempting to make concerned the issue of the relationship between the formal string of symbols that constitute "G(F)" and "Omega(F)" and the actual *meanings* that these strings are supposed to represent. I was merely trying to argue that meanings are essential - a point with which Feferman strongly concurs, in his commentary.) It should be made clear that none of these corrections affects the arguments of Chapter 3 in any way (so long as Omega(F) is replaced by G(F) throughout), as Feferman himself appears to affirm in his last paragraph of the aforementioned section.

2.5 I find it unfortunate, however, that he does not offer any critique of the arguments of Chapter 3. I would have found it very valuable to have had the comments of a first-rate logician such as himself on some of the specifics of the discussions in Chapter 3. Feferman seems to be led to having some unease about the arguments presented there, not because of specific errors that he has detected, but merely because my "slapdash scholarship" may be "stretched perilously thin in areas different from [my] own expertise". A related point is made by McCarthy, McDermott and Baars in connection with my evidently inadequate referencing of the literature on AI, and on other theories that relate to consciousness, either in its computational, biological, or psychological respects.

2.6 I think that a few words of explanation, from my own vantage point, are necessary here. An ability to search thoroughly through the literature has never been one of my strong points, even in my own subject (whatever that might be!). My method of working has tended to be that I would gather some key points from the work of others and then spend most of my time working entirely on my own. Only at a much later stage would I return to the literature to see how my evolved views might relate to those of others, and in what respects I had been anticipated or perhaps contradicted. Inevitably I shall miss things and get some things wrong. The most likely source of error tends to be with second-hand information, where I might misunderstand what someone else tells me when

reporting on the work of a third person. Gradually these things sort themselves out, but it takes time.

2.7 My reason for mentioning this is to emphasize that errors of the nature of those pointed out by Feferman are concerned essentially with this link of communication with the outside (scientific, philosophical, mathematical, etc.) world, and not with the internal reasonings that constitute the essential Gödelian arguments of *Shadows*. Most specifically, the main parts of Chapter 3 (particularly 3.2, 3.3 and 3.5-3.24) are entirely arguments that I thought through on my own, and are therefore independent of however "slapdash" my scholarship might happen to be! I trust that these arguments will be judged entirely on their

intrinsic merits.

# 3. The Central New Argument of *Shadows*

3.1 Chalmers provides a succinct summary of the central new argument that I presented in *Shadows* (Section 3.16, and also 3.23 and 3.24 - but recall that my Omega(F) should be replaced by G(F) throughout Section 3.16 and 3.23). Let me repeat the essentials of Chalmers's presentation here - but with one important distinction, the significance of which I shall explain in a moment.

3.2 We try to suppose that the totality of methods of (unassailable) mathematical reasoning that are in principle humanly accessible can be encapsulated in some (not necessarily computational) sound formal system F. A human mathematician, if presented with F, could argue as follows (bearing in mind that the phrase "I am F" is merely a shorthand for "F encapsulates all the humanly accessible methods of mathematical proof"):

(A) "Though I don't know that I necessarily am F, I conclude that if I were, then the system F would have to be sound and, more to the point, F' would have to be sound, where F' is F supplemented by the further assertion "I am F". I perceive that it follows from the assumption that I am F that the Gödel statement G(F') would have to be true and, furthermore, that it would not be a consequence of F'. But I have just perceived that "if I happened to be F, then G(F') would have to be true", and perceptions of this nature would be precisely what F' is supposed to achieve. Since I am therefore capable of perceiving something beyond the powers of F', I deduce that, I cannot be F after all. Moreover, this applies to any other (Gödelizable) system, in place of F."

3.3 (Of course, one might worry about how an assertion like "I am F" might be made use of in a logical formal system. In effect, this is discussed with some care in *Shadows*, Sections 3.16 and 3.24, in relation to the Sections leading up to 3.16, although the mode of presentation there is somewhat different from that given above, and less succinct.)

3.4 The essential distinction between the above presentation and that of Chalmers is that he makes use (in the first (2) of his Section 2) of the stronger conditional assumption "I know that I am F", rather than merely "I am F", the latter being all that I need for the above. Thus, if we accept the validity of the above argument, the conclusion is considerably stronger than the "strong" conclusion that Chalmers draws ("threatening to the prospects of AI") to the effect that it "would rule out even the possibility that we could empirically discover that we were identical to some system F".

3.5 In fact, it *was* this stronger version (A) that I presented in *Shadows*, from which we would conclude that we cannot be identical to any knowable (Gödelizable) system F whatever, whether we might empirically come to believe in it or not! I am sure that this stronger conclusion would provide an even greater motivation for people (whether AI supporters or not) to find a flaw in the argument. So let me address the particular objection that Chalmers (and, in effect, also McCullough) raises against it.

3.6 The problem, according to Chalmers, is that it is contradictory to "know that we are sound". Accordingly, he argues, it would be invalid to deduce the soundness of F, let alone that of F', from the assumption "I am F". On the face of it, to a mathematician, this seems an unlikely let-out, since in all the above discussions we are referring simply to the notion of *mathematical proof*. Moreover, the "I" in the above discussion refers to an idealized human mathematician. (The problems that this notion raises, such as those referred to by McDermott, are not my concern at the moment. I shall return to such matters later; cf. Section 6.) Suppose that F indeed represents the totality of the procedures of mathematical proof that are in principle humanly accessible. Suppose, also, that we happen to come across F and actually entertain this possibility that we might "be" F, in this sense (without actually knowing, for sure, whether or not we are indeed F). Then, under the assumption that it *is* F that encapsulates all the procedures of valid mathematical proof, we must surely conclude that F is sound. The whole point of the procedures of mathematical proof is that they instil *belief*. And the whole point of the Gödel argument, as I have been employing it, is that a belief in the conclusions that can be obtained using some system H entails, also, a belief in the soundness and consistency of that system, together with a belief (for a Gödelizable H) that this consistency cannot be derived using H alone.

3.7 This notwithstanding, Chalmers and McCullough argue for an inconsistency of the very notion of a "belief system" (which, as I have pointed out above, simply means a system of procedures for mathematical proof) which can believe in itself (which means that mathematicians actually trust their proof procedures). In fact, this conclusion of inconsistency is far too drastic, as I shall show in a moment. The key issue is not that belief systems are inconsistent, or incapable of trusting themselves, but that they must be restricted as to what kind of assertion they are competent to address.

3.8 To show that "a belief system which believes in itself" need not be inconsistent,

consider the following. We shall be concerned just with P-sentences (which, we recall, are assertions that specified Turing machine actions do not halt). The belief system B, in question, is simply the one which "believes" (and is prepared to assert as "unassailably perceived") a P-sentence S if and only if S happens to be true. B is not allowed to "output" anything other than a decision as to whether or not a suggested P-sentence is true or false - or else it may prattle on, as is its whim, generating P-sentences together with their correct truth values. However, as part of its internal musings, it is allowed to contemplate other kinds of thing, such as the fact that it *does*, indeed, produce only truths in its decisions about P-sentences. Of course, B is not a computational system - it is a Turing *oracle* system, as far as its output is concerned - but that should not matter to the argument. Is there anything wrong in B "believing in the soundness of B"? Nothing whatever, if we interpret this in the right way. The important thing is that B is allowed only to make assertions about P-sentences. It can use whatever procedures it likes in its internal musings, but all its *outputs* must be assertions as to the validity of particular P-sentences. If we apply the diagonal procedure that Chalmers and McCullough refer to, then we get something which is not a P-sentence, and is accordingly not allowed to be part of this belief system's output.

3.9 It may be felt that this is a pretty limited kind of "belief system", where it can make assertions only about the truth or falsity of P-sentences. Perhaps it is limited; but it is precisely a belief system of this very kind that comes into the arguments of Chapter 3 of *Shadows*. In that discussion, I was careful, in the key Section 3.16 of *Shadows*, to limit the mathematical assertions under consideration to P-sentences. This avoids many difficult issues that can arise without such restrictions. However, the robots described in that section are allowed to think in very general terms - as human mathematicians may do - about non-computable systems and uncountable cardinals, etc. Nevertheless, the *-assertions under consideration must always be P-sentences, and it is *only* in relation to such sentences (as outputs) that the formal systems Q(M) and Q_M(M) are constructed. In this circumstance the argument serves to show that the robots' "belief system" cannot, after all, be a computational one, provided that it is broad enough to encompass Gödelian reasoning - which is a contradiction with the notion of "robot" that was being used.

3.10 This is not to say that the diagonalization procedure that McCullough and Chalmers refer to need apply only to computational belief systems F. As they both argue (particularly McCullough), there is no requirement that F be computational in their discussions. Indeed, in Section 7.9 of *Shadows* (which is in Part 2, so it is easy to miss, if one is concerned only with the logical arguments of that book - and neither McCullough nor Chalmers actually mention it), I explicitly referred to the fact that the Gödel-type diagonalization arguments of Part 1 will apply much more generally than merely to computational systems. For example, if Turing's oracle-computation notions are adopted, then the diagonalization procedures are quite straight-forward. However, in any specific application, it is necessary to restrict the class of sentences to which the notion of "unassailable belief" can be applied. If we do not do this, we can land in paradox, which is exactlythe situation that McCullough and Chalmers find themselves in.

3.11 Indeed, McCullough actually carries through such paradoxical reasoning in his Section 2.1, seeming to be presenting this parody of my own reasoning as though it were actually my own reasoning. This is beneath his usual standards. It would have been more helpful if he had addressed the arguments as I actually presented them.

3.12 Returning to the argument (A), we now see how to avoid the inherent difficulties that occur with a belief system with an unrestricted domain. A sufficient thing to do is to make sure that the word "sound" is interpreted in the restricted sense which applies only to P-sentences - as was indeed done in *Shadows*, Section 3.16. (Recall the discussion of Section 2, above, in which Feferman draws attention to possible differences of interpretation of that word.) This provides the needed argument against computationalism, and it is not subject to the objection brought forward by Chalmers in his discussion of my "second argument" in his Section 2.

3.13 Of course, as in Section 7.9 of *Shadows* and as in McCullough's discussion, it is possible to repeat this argument at a higher level. Rather than restricting attention to P-sentences (that is, PI_1 sentences), we could use PI_2-sentences, say (cf. Feferman's commentary). The diagonal process can be applied, but it does not yield a PI_2-sentence, so contradiction (of the Chalmers/McCullough type - to a self-believing belief system) is again avoided. The same argument applies to higher-order sentences. However, it is important to put *some* restriction on the type of sentence to which the belief system is applied. This kind of thing is very familiar in mathematical logic. One may reason about sets, and about sets of sets, and sets of sets of sets, etc., but one cannot reliably reason about the set of *all* sets. That leads immediately to a contradiction, as Cantor and Russell pointed out long ago. Likewise, a self-believing belief system cannot consistently operate if it is allowed to apply itself to unrestricted mathematical systems. In Section 3.24 of *Shadows*, I tried to explore the tantalizing closeness that my Gödelian reasoning of Section 3.16 seemed at first to have with the Russell-type reasoning that leads to paradox. My conclusion was that the argument of Section 3.16, as I presented it, was not actually of the same nature at all, since the domain of consideration (P-sentences) was indeed sufficiently restricted. I am well aware that the argument can be taken much further than this, and it would be interesting to know how far. Moreover, it would be interesting to have a professional logician's commentary on all these lines of thinking.

# 4. The "Bare" Gödelian Case

4.1 Although I have concentrated, in the previous section, on what I have referred to as the "central new argument" of *Shadows*, I do not regard that as the "real" Gödelian reason for disbelieving that computationalism could ever provide an explanation for the mind - or even for the behaviour of a conscious brain.

4.2 Perhaps a little bit of personal history on this point would not be amiss. I first heard about the details of Gödel's theorem as part of a course on mathematical logic (from

which I also learned about Turing machines) given by the Cambridge logician Steen. As far as I can recall, I was in my first year as a graduate student (studying algebraic geometry) at Cambridge University in 1952/53, and was merely sitting in on the course as a matter of general education (as I did with courses in quantum mechanics by Dirac and general relativity by Bondi). I had vaguely heard of Gödel's theorem prior to that time, and had been a little unsettled by my impressions of it. My viewpoint before that would probably have been rather close to what we now call "strong AI". However, I had been disturbed by the possibility that there might be true mathematical propositions that were in principle inaccessible to human reason. Upon learning the true form of Gödel's theorem (in the way that Steen presented it), I was enormously gratified to hear that it asserted no such thing; for it established, instead, that the powers of human reason could not be limited to any accepted preassigned system of formalized rules. What Gödel showed was how to transcend any such system of rules, so long as those rules could themselves be trusted.

4.3 In addition to that, there was a definite close relationship between the notion of a formal system and Turing's notion of effective computability. This was sufficient for me. Clearly, human thought and human understanding must be something beyond computation. Nevertheless, I remained a strong believer in scientific method and scientific realism. I must have found some reconciliation at the time which was close to my present views - in spirit if not in detail.

4.4 My reason for presenting this bit of personal history is that I wanted to demonstrate that even the "weak" form of the Gödel argument was already strong enough to turn at least one strong-AI supporter away from computationalism. It was not a question of looking for support for a previously held "mystical" standpoint. (You could not have asked for a more rationalistic atheistic anti-mystic than myself at that time!) But the very force of Gödel's logic was sufficient to turn me from the computational standpoint with regard not only to human mentality, but also to the very workings of the physical universe.

4.5 The many arguments that computationalists and other people have presented for wriggling around Gödel's original argument have become known to me only comparatively recently: perhaps we act and perceive according to an unknowable algorithm; perhaps our mathematical understanding is intrinsically unsound; perhaps we could know the algorithms according to which we understand mathematics, but are incapable of knowing the actual roles that these algorithms play. All right, these are logical possibilities. But are they really plausible explanations?

4.6 For those who are wedded to computationalism, explanations of this nature may indeed seem plausible. But why *should* we be wedded to computationalism? I do not know why so many people seem to be. Yet, some apparently hold to such a view with almost religious fervour. (Indeed, they may often resort to unreasonable rudeness when they feel this position to be threatened!) Perhaps computationalism can indeed explain the

facts of human mentality - but perhaps it cannot. It is a matter for dispassionate discussion, and certainly not for abuse!

4.7 I find it curious, also, that even those who argue dispassionately may take for granted that computationalism in some form - at least for the workings of the objective physical universe - *has* to be correct. Accordingly, any argument which seems to show otherwise *must* have a "flaw" in it. Even Chalmers, in his carefully reasoned commentary, seeks out "the deepest flaw in the Gödelian arguments". There seems to be the presumption that whatever form of the argument is presented, it just *has* to be flawed. Very few people seem to take seriously the slightest possibility that the argument might perhaps, at root, be correct! This I certainly find puzzling.

4.8 Nevertheless, I know of many who (like myself) do find the simple "bare" form of the Gödelian argument to be very persuasive. To such people, the long and sometimes tortuous arguments that I provided in *Shadows* may not add much to the case - in fact, some have told me that they think that they effectively weaken it! It might seem that if I need to go to lengths such as that, the case must surely be a flimsy one. (Even Feferman, from his own particular non-computational standpoint, argues that my diligent efforts may be "largely wasted".) Yet, I would claim that some progress has been made. I am struck by the fact that none of the present commentators has chosen to dispute my conclusion G (in *Shadows*, p.76) that "Human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth". I doubt that any will admit to being persuaded by any of the replies to my queries Q1, ..., Q20, in Section 2.6 and Section 2.10, but it should be remarked that many of these queries represented precisely the kinds of misunderstandings and objections that people had raised against my earlier use of the bare Gödelian argument (and its conclusion G) in *The Emperor's New Mind*, particularly in the many commentaries on that book in *Behavioral and Brain Sciences* (and, in particular, one by McDermott 1990). Perhaps some progress has been made after all!

# 5. Gödel's "Theorem-Proving Machine"

5.1 Before addressing the important issue of possible errors in human reasoning or the possible "unknowability" of the putative algorithm underlying human mathematical reasoning (which provide the counter-arguments that so many computationalists pin faith on), I should briefly refer to the discussion of Section 3.3 in *Shadows*, which Chalmers regards as "one of the least convincing sections in the book". This is the first of the two arguments of mine that he addresses, but I am not sure that he (or any other of the commentators) has appreciated what I was trying to express. In that section (and also Section 3.8, cf. figure 3.1 on p. 148), I was attempting to show the actual absurdity of the possibility that human understanding (with regard to P-sentences, say) might be encapsulated in what I have referred to as a "Gödel's theorem-proving machine". As quoted on p. 128 of *Shadows*, Gödel seemed not to have been able to rule out the possibility that mathematical understanding might be encapsulated in terms of the action of an algorithm - his "theorem-proving machine" - which, although sound, could not be

humanly (unassailably) perceived to be sound. Yet it might be possible to come across this algorithm empirically. I shall refer to this putative "machine" (or algorithm) here as T.

5.2 In Section 3.3, I was concerned with a mathematical algorithm, of the type that might be considered seriously by logicians or mathematicians, so it is not unreasonable to think of T as formulated in the kind of terms which mathematical logicians are familiar with. Of course, even if T were not initially formulated in such terms, it could be if desired. It is sufficient to restrict Gödel's hypothetical theorem-proving machine to be concerned only with P-sentences. Then T would be an algorithmic procedure that generates precisely all the true P-sentences that are perceivably true, in principle, by human mathematicians. Gödel argues that although T might be empirically discoverable, the perception of its soundness would be beyond the powers of human insight. In Sections 3.3 and 3.8, I merely try to make the case that the existence of T is a very far-fetched possibility indeed, especially if we try to imagine how it might have come about (either by natural selection or by deliberate AI construction). But I did not argue that it was an entirely illogical possibility.

5.3 In Feferman's commentary, he refers to Boolos's "cautious" interpretation of the implications of Gödel's theorem that a let-out for computationalism would be the existence of "absolutely unsolvable diophantine problems". Such an absolutely unsolvable problem could be constructed, by well understood procedures, from the algorithm T, if T were to exist. Phrased in these terms, it does not seem at all out of the question that such a T might exist. In Section 3.3, my intention was merely to point out some of the improbable-sounding implications of the existence of T. It seems to me that this does go somewhat beyond what Feferman refers to at the end of his commentary. Moreover, the arguments referred to in Section 2 above (concerning Section 3.16 of *Shadows* that most commentators appear to have missed) certainly do proceed well beyond this interpretation.

5.4 Later in *Shadows* (cf. Sections 3.5-3.23, and especially 3.8), I argue that it is extremely hard to see how an extraordinarily sophisticated algorithm of the nature of T could come about by natural selection (or by deliberate AI construction), even if it could exist. It has to be already capable of correctly dealing with subtle mathematical issues that are, for example, far beyond the capabilities of the Zermelo-Fraenkel axiom system ZF (for example, the Gödel procedure can be applied to ZF to obtain humanly accessible P-sentences that are indeed beyond the scope of ZF). Yet issues of this nature played no role in the selective processes that were operative with our remote ancestors. I would argue that there is nothing wrong with natural selection having been the driving force, so long as it is a non-specific *non-computational* quality such as "understanding" that natural selection has favoured, rather than some improbable algorithm, such as T.<3>

5.5 Even if we do not worry about how T might possibly have come about, there is a

distinct implausibility in its very existence, if T were to be an algorithm that could be humanly understood (or "knowable", in the terminology of *Shadows*). This is basically "case II" of *Shadows* (cf. p. 131), where the soundness of T, and certainly its specific role, would not be humanly knowable. The implausibility of such a T was the main point that I was trying to make in Section 3.3. I think Chalmers is arguing that such a T might come about by some bottom-up AI procedures and, if so, it might not look at all like a mathematical formal system. However, in the absence of some strongly held computationalist belief - to the effect that it *must* have been by procedures of this very kind that Nature was able to produce human mathematicians - there is no good reason to expect that this would be a good way of finding such a T (as I argue in *Shadows* Section 3.27), nor is there any reason to expect such a T actually to exist. It was the burden of later sections of *Shadows*, not of Section 3.3, to argue that such bottom-up procedures do not do what is required either. In effect, in these later sections, I argue that if merely the (partly bottom-up) computational *mechanisms* for ultimately leading to a T could be known, then we would indeed be able to construct the formal system that T represents. This will be discussed further in Section 7, below.

# 6. The Issue of Errors

6.1 Some commentators (particularly McDermott and, in effect, Baars) try to argue that the fact that human mathematicians make errors allows the computational model of the mind to escape the Gödel-type arguments. (This was also apparently Turing's let-out, as illustrated in the quote in *Shadows*, p.129.) I have stressed in many places in *Shadows* that the main arguments of that book (certainly those in Chapter 2) are concerned with what mathematicians are able to perceive *in principle*, by their methods of mathematical proof - and that these methods need not be necessarily constrained to operate within the confines of some preassigned formal system.

6.2 I fully accept that individual mathematicians can frequently make errors, as do human beings in many other activities of their lives. This is not the point. Mathematical errors are in principle correctable, and I was concerned mainly with the ideal of what can indeed be perceived in principle by mathematical understanding and insight. Most particularly, I was concerned with those P-sentences that can be humanly perceived, in principle, i.e., with those which are in principle humanly accessible. The arguments given above, in Sections 3 and 5, were also concerned with this ideal notion only. The position that I have been strongly arguing for is that this ideal notion of human mathematical understanding is something beyond computation.

6.3 Of course, individual mathematicians may well not accord at all closely with this ideal. Even the mathematical community as a whole may significantly fall short of it. We must ask whether it is conceivable that this mathematical community, or its individual members, could be entirely computational entities even though the ideal for which they strive is beyond computation. Put in this way, it may perhaps seem not unreasonable that this could be the case. However, there remains the problem of what the human

mathematicians are indeed doing when they seem able to "strive for", and thereby approximate, this non-computational ideal. It is the *abstract idea* underlying a line of proof that they seem able to perceive. They then try to express these abstract notions in terms of symbols that can be written on a page. But the particular collections of symbols that ultimately appear on the pages of their notes and articles are far less important than are the ideas themselves. Often the particular symbols used are quite arbitrary. With time, both the ideas and the symbols describing them may become refined and sometimes corrected. It may not always be very easy to reconstruct the ideas from the symbols, but it is the *ideas* that the mathematicians are really concerned with. These are the basic ingredients that they employ in their search for idealized mathematical proofs. (These matters have relevance to the question of how mathematicians *actually* think,<4> as raised by Feferman in his commentary, and they are related also to issues raised also by Baars and McCullough.)

6.4 Sometimes there may be errors, but the errors are correctable. What is important is the fact is that there *is* an impersonal (ideal) standard against which the errors can be measured. Human mathematicians have capabilities for perceiving this standard and they can normally tell, given enough time and perseverance, whether their arguments are indeed correct. How is it, if they themselves are mere computational entities, that they seem to have access to these non-computational ideal concepts? Indeed, the ultimate criterion as to mathematical correctness is measured in relation to this ideal. And it is an ideal that seems to require use of their conscious minds in order for them to relate to it.

6.5 However, some AI proponents seem to argue against the very existence of such an ideal, a position that Moravec (if his robot is to be trusted as espousing Moravec's own views) seems to be taking in his commentary. Moreover, Chalmers comments: "an advocate of AI might take [the position] that our reasoning is fundamentally unsound, even in idealization". There are others, such as Baars ("I do not believe in the absolute nature of mathematical thought"), who also have difficulty with this notion, perhaps because their professional interests have more to do with examining the ways in which particular individuals may deviate from such ideals than with the ideal notions themselves. It is common for such people to point to errors that have persisted in the mathematical literature for some while (such as McDermott 's reference to Kempe's erroneous attempt at a proof of the four-colour theorem - which, incidentally provided an important ingredient in the actual proof that was finally arrived at in 1976 by Appel and Haken; cf. Devlin (1988) - or to Frege's inconsistent attempt at building up a formal set theory - which was a good deal more influential, in a very positive sense). But these errors are more in the nature of "correctable errors", and do not really argue against the very existence of a mathematical ideal.

6.6 In *Shadows*, Section 3.2, I did examine, in a serious way, the possibility that mathematical reasoning might be fundamentally unsound. But one should bear in mind that the presumption of mathematical unsoundness is an extremely dangerous position for anyone purporting to be a scientist to take. If our mathematical reasoning were indeed

fundamentally unsound, then the whole edifice of scientific understanding would come crashing to the ground! For virtually all of science, at least detailed science, depends upon mathematics in one respect or another. I find it remarkable how frequently attacks on the Gödelian argument seem to degenerate into attacks upon the very basis of mathematics.<5> To attack the notion of "ideal" mathematical concepts or idealized mathematical reasoning is, indeed, to attack the very basis of mathematics. People who do so should at least pause to contemplate the implications of what they are contending.

6.7 While it is true that there are different philosophical standpoints that may be adopted by different mathematicians, this has little effect on the basic Gödelian argument, especially if we restrict attention to P-sentences; see responses to queries Q9-Q13 in Sections 3.6, 3.10 of *Shadows*. For the remainder of my arguments here, I shall take it as read that there *is* an ideal notion of (in principle) humanly accessible mathematical proof, at least with respect to P-sentences, and that this ideal notion of proof is *sound*. (And I am not against there being *more* than one, provided that they are not in contradiction with one another with regard to P-sentences; see *Shadows* Section 3.10, response to Q11.) The question, then, is how serious are the errors which undoubtedly occur when actual human mathematicians attempt toemulate this ideal.

6.8 For the arguments of Chapter 3 of *Shadows*, particularly Sections 3.4, 3.17, 3.19, 3.20, and 3.21, I try to address the issue of errors in purported mathematical arguments, and the question of constructing an error-free formal system from the *actual* output of a manifestly computational system - the hypothetical mathematical robots that I consider for the purpose. The arguments are quite intricate in places, and I do not blame some of the commentators for balking at those sections. On the other hand, it would have been helpful to have had a dispassionate discussion of these arguments in their essential points. McDermott does at least address some of the more technical arguments concerning errors - though I feel it is not altogether appropriate to refer to his account as "dispassionate". More importantly, he does not answer the essential point of my conclusions. If it is to be *errors* that provide the key escape route from the Gödel conundrum, we need to explain the seeming necessity for a "conspiracy" preventing any kind of computational procedure for weeding out all the errors in the merely *finite* set that arises in accordance of the discussion of Section 3.20 (see 3.21 and also the second paragraph of 3.28). In his commentary McDermott does not actually address the argument as I gave it, but goes off on a tangent (about a "computerized Gauss" and the like) which has very little to do with the specific argument provided in *Shadows*. (The same applies to most of his other arguments which, he contends, have "torn [my] argument to shreds". His discussion might have been more convincing had it referred to my actual arguments! I shall make some further comments concerning these matters in Section 7 below.)

6.9 McDermott does, however, come close to expressing the central dilemma presented by the Gödelian insight - although apparently unwittingly. He has a hard time coming to terms with the fact that mathematical unassailability needs "to be both informal and guaranteed accurate". Although he is unable to "see how that's possible", it is basically

this conflict that forces us into a non-computational viewpoint. If by a "guaranteed accurate" notion of unassailability he means something that has been validated by a procedure that is computationally checkable, then this notion would basically have to be one that can indeed be encompassed by a formal system in the ordinary sense. We must bear in mind that the guarantee must apply not only to the correctness of *carrying out* the rules of the procedure (which is where the "computational checkability" of the procedure might have importance), but also to the validity, or *soundness* of the very rules themselves. But if we can guarantee that the rules are sound, we can also guarantee something beyond those rules. The rules would be subject to Gödel's theorem, so there would also be certain P-sentences, such as the Gödel sentence asserting the consistency of the "guaranteeing system", that would be just as "guaranteed" as the things that have already been previously "guaranteed". If McDermott is requiring that "formal" implies "computational", and that "guaranteeable" also implies computational, then he has a problem encompassing certain things that mathematicians are actually capable of guaranteeing, namely the passing from a given guaranteeing system to the implied guaranteeing of its Gödel sentence.

6.10 One of the key points of the discussion of Chapter 3 of *Shadows* was to exhibit the importance of this conflict within the context of an entirely computational framework. If we accept that the putative robots described there are entirely computational entities, then any "guaranteeing" system that they come up with must necessarily be computational also. Accepting that the robots must also guarantee their guaranteeing system (see Section 3 above) and that they appreciate Gödel's theorem - and also accepting that random elements play no fundamentally important role in their behaviour (see 3.18, 3.22) - we are driven to the remaining loophole for computationalism: errors. It was the thrust of Sections 3.17-3.21 to demonstrate the implausibility of this loophole also. For this discussion, one attempts to find computationally bounded safeguards against errors, and then shows that this is impossible.

6.11 In effect, though in a stronger form than usual, all this is saying is that it is impossible to "formalize" the informal notion of unassailable mathematical demonstration. In this sense McDermott is indeed right to fail to "see how that's possible". It's *not* possible if "formalize" indeed implies something *computational*. That's the whole point!

# 7. The "Unknowability" Issue

7.1 Several other commentators (Chalmers, Maudlin, Moravec - and also McDermott again!) prefer to attack the Gödel argument from the standpoint that the "algorithm" (or formal system) to which Gödel's theorem is to be applied is *unknowable* in some sense - or, at least, unknowable to the person attempting to apply the argument. (Indeed, Chalmers, for one, seems to be happy enough to accept "that we have an underlying sound competence, even if our performance sometimes goes astray"; so in his

commentary on my "First Argument" - that given in *Shadows*, Section 3.3 - he seems to be resorting to the "unknowability" of the algorithm in question.)

7.2 There is an unfortunate tendency for some people (Chalmers, and some others excepted) to try to twist my use the Gödel argument away from the form in which I actually gave it, which refers to "mathematical understanding" in the abstract sense - or at least in the sense in which that term might apply to the mathematical community as a whole - to a more personal form. Such people seem to regard it as more impressively ridiculous that some individual mathematician could know his or her "personal algorithm", than that the principles underlying the proof procedures that are common to mathematicians as a whole might be accessible to the common understanding of the mathematical community. And they apparently regard it as particularly evidently ridiculous that I myself should have such access (cf. commentaries by McCullough, Maudlin, and Moravec), so they phrase what they take to be my own Gödelian arguments in the form of what kind of a contradiction I might land myself in if I happened to come across my own personal algorithm! I suppose that in order to make "debating points", such procedures may seem effective, but I find it distinctly unhelpful to phrase the arguments in this way; for the arguments then become significantly changed from the ones that I actually put forward.

7.3 Particularly unhelpful are formulations like Moravec's "Penrose must err to believe this sentence." and McCullough's "This sentence is not an unassailable belief of Roger Penrose." Although there are ways of appreciating the nature of the particular sentence that Gödel originally put forward in terms that are not totally dissimilar from this, it is certainly a travesty to attempt to express the essentials of my own (or indeed Gödel's) argument in this way. Only marginally better would be "No mathematician can believe unassailably that this sentence is true." or "No conscious being can accept the truth of this sentence." - mainly because of their manifest similarity to the archetypal self-contradictory assertion: "This sentence is false." In Section 3.24 of *Shadows*, I explicitly addressed the possibility that the kind of reasoning that I had been using earlier in the book (basically the argument of 3.16, which is that of Section 3 above, but also 3.14) might be intrinsically self-contradictory in this kind of way. I do not think that it is, for reasons that I discussed in 3.24. None of the commentators has chosen to dispute me on this particular issue, so perhaps I may take it that they agree also!

7.4 Instead, the arguments, relevant to the present discussion, that Chalmers, Maudlin, McDermott, and Moravec are really putting forward (and which are greatly obscured by the above kind of formulation), is that the algorithm in question might be *unknowable*. They make the point that in order to provide an effective simulation of the thought processes of an individual mathematician, an almost unimaginably complicated algorithm would have to be envisaged. Of course, this point had not escaped me either(!), which is the main reason why I formulated my own discussion in quite different ways from this.

7.5 There are, in fact, two distinct broad lines of argument put forward in *Shadows*, the *simple* argument and the *complicated* argument. The simple argument (which has always

been good enough for me) is basically the "bare" Gödelian reasoning referred to in Section 4 above (leading to the conclusion G of *Shadows*, p.76), as applied to the mathematicians' belief that they are "really doing what they think they are doing", rather than blindly following the rules of some unfathomable algorithm (see the opening discussion of 3.1 and the final one of 3.8). Accordingly, the procedures available to mathematicians *ought* all to be knowable! The only remark concerning any aspect of implications of this line of argument that I can find in these commentaries is that towards the end of McDermott's piece, in which he remarks that the quality of conscious understanding will, in his view, turn out to be something "quite simple" (because "consciousness is no big deal"). I remarked (*Shadows*, p.150) that "understanding has the appearance of being a simple and common-sense quality", but if it actually is something simple, it has to be something non-computational, because otherwise it becomes subject to the bare form of the Gödelian argument. I do not think that McDermott would be very happy with that, but he does not refer to this particular problem. (As an aside, I find it hard to see why some commentators, such as Maudlin, seem to argue that the slightest flaw in the discussion of Part 1 of *Shadows* would demolish the whole argument. In fact there are several different lines of argument presented there. All of them would have to be demolished independently!)

7.6 The complicated lines of argument are addressed more at those who take the view that mathematicians are not "really doing what they think they are doing", but are acting according to some unconscious unfathomable algorithm. Since there is no way that we could know what this algorithm is (or what several distinct but effectively equivalent algorithms might be; cf. *Shadows*, Section 3.7), I adopt a completely different line of approach. This is to examine how such an unfathomable algorithm might conceivably come about. The issue of the role of natural selection (treated particularly in 3.8) was referred to in Section 5 above. The other possibility that I discussed was some form of deliberate AI construction, and that was the thrust of Sections 3.9 onwards, in Chapter 3 of *Shadows*.

7.7 Rather than trying to "know" whatever putative algorithm might now describe the physical action of the brain of some individual human mathematician - or else what complicated computer program might now control the actions of some putative intelligent mathematics-performing individual robot - I consider the general type of computational AI process that might underlie the evolution of such a robot. We do not need to know how the robot's computer-brain is actually supposed now to be wired up, since I am prepared to accept that the "bottom-up" procedures that are used (artificial neural networks, genetic algorithms, random inputs, even natural selection processes that might be applied to the robots themselves, etc. - and also adequately simulated environments, cf. *Shadows* Section 3.10, McCarthy, McCullough, and McDermott take note) could lead to a final product of almost unimaginable complication. Nevertheless, these very *mechanisms* that go into the ultimate construction of the robots would indeed be knowable - in fact, it might well be claimed (as I know that Moravec (1988) has actually claimed) that these mechanisms are, in effect, known already. The whole point of considering these mechanisms, rather than the "actual" algorithm that is supposed to be enacted by the computer-brain of our putative robot (a point apparently missed by

Maudlin, McDermott, and Moravec), is that the former would be supposed to be knowable, so long as those aspects of the AI programme that are aimed at the construction of an actually intelligent robot - intelligent enough for it to be able to understand mathematics - are attainable within the general framework of present-day computer-driven ideas.

# 8. AI and MJC

8.1 A summary of this line of reasoning formed part of the "fantasy dialogue" given in Section 3.23 of *Shadows*. (In what follows, "MJC" refers to the robot, whereas "AI" refers to the subject of artificial intelligence.) Thus, when Maudlin ridicules the possibility that MJC might "easily 'digest' its own algorithm", he has missed the point. There is not supposed to be anything "unknowable" about the procedures of AI; otherwise there would be no point in people actually trying to *do* AI!

8.2 One of the aims of the discussion in the dialogue was to bring home the fact that, according to the "optimistic" school of AI, to which Moravec belongs, it need not be so far into the future when robots are actually constructed which could exceed all human mental capabilities. In particular, such a robot could perform feats of mathematical understanding that exceed those of any human mathematician. This indeed seems to be a corollary of such an optimistic stance with regard to AI, and is not particularly (as McDermott contends) "extravagant" from the point of view of the tenets of AI. My characterization of MJC was to set its abilities, with regard to mathematical understanding, just ahead of that of humanity, but with a particularly effective ability with regard to directly computational matters. Thus, it would have no difficulty at all in assimilating the purely computational aspects of the mechanisms concerning its original construction (since these were in any case already known to Albert Imperator, but they might be computationally very involved), whilst MJC might be relatively slower in appreciating the subtleties of certain logical points - although still a good deal faster than one might imagine a human mathematician would be.

8.3 This does not seem to me to be an "incongruity" in characterization, as Moravec seems to suggest. Of course MJC goes mad at the end - but why not? It has just been driven to the logical conclusion that the only way in which it could have come about was by God implanting a Divine Algorithm into its mechanisms, through the "chance" elements that were part of those mechanisms. It is not a question of MJC suddenly realizing that its initials stand for "Messiah Jesus Christ", as McDermott seems to think. (The initials were just intended as a little joke for the reader, and not really part of the story.) In fact, McDermott seems extraordinarily slow in getting to the point of the story, if indeed he ever really gets to the point of it. (Actually, it seems that he does not, especially in view of his comments about "affixing a * to Omega(Q)", etc. He has not appreciated the central argument repeated in Section 3 above. It is clear that Maudlin misses the point also, since the dialogue has nothing to do with "a computer failing to pass the Turing test". But so also does Moravec's robot, so McDermott and Maudlin are

in good company!) I certainly do *not* believe that a computationally controlled robot could achieve the kind of easy-flowing intelligent-sounding dialogue that MJC exhibits. That is the whole point of a reductio ad absurdum. One assumes that all the implications of the premise, that one intends ultimately to disprove, actually hold good. The final contradiction disproves the premise. Here, the premise is that the procedures of computational AI can ultimately lead to the construction of an intelligent mathematics-performing robot. Of course such a putative robot could be articulate and sound intelligent in other ways than just in mathematics. But it doesn't mean that I believe the premise.

8.4 On another point, the fantasy dialogue does not actually summarize *all* the arguments of Chapter 3 of *Shadows*. Most particularly, it does not summarize most of the arguments given in Sections 3.17-3.21 against the "errors" argument (cf. Section 6 above). I did not include these mainly because I felt that the discussion was already getting rather long and complicated; and since the "error" discussion was rather involved, I thought it best to leave most of it out. In addition to this, the way that the dialogue developed, it seemed appropriate for MJC to have a distinctly arrogant character. It would have changed the flavour of the story to allow MJC to acquire the humility that would have been needed in order to have it admit to being subject to serious error.

8.5 In some ways this was perhaps unfortunate, because it appears to give Maudlin and McDermott an easy way out by allowing their allegedly more "realistic" version of MJC to make the occasional mistake. This, however, would be to miss the point of the "errors" arguments, as given in *Shadows* and as referred to in Section 6 above.

# 9. Mathematical Platonism

9.1 I think that a few remarks in relation to my attitude to mathematical Platonism are appropriate at this stage. Indeed, certain aspects of my discussion of errors, as given in Section 6 above, might seem to some to be inappropriately "Platonistic", as they refer to idealized mathematical arguments as though they have some kind of existence independently of the thoughts of any particular mathematician. However, it is difficult to see how to discuss abstract concepts in any other way. Mathematical proofs are concerned with abstract ideas - ideas which can be conveyed from one person to another, and which are not specific to any one individual. All that I require is that it should make sense to speak of such "ideas" as real things (though not in themselves material things), independent of any particular concrete realization that some individual might happen to find convenient for them. This need not presuppose any very strong commitment to a "Platonistic" type of philosophy.

9.2 Moreover, in the particular Gödelian arguments that are needed for Part 1 of *Shadows*, there is no need to consider as "unassailable", any mathematical proposition other than a P-sentence (or perhaps the negation of such a sentence). Even in the very

weakest form of Platonism, the truth or falsity of P-sentences is an absolute matter. I should be surprised if even Moravec's robot could make much of a case for alternative attitudes with regard to P-sentences (though it is true that some strong intuitionists have troubles with unproved P-sentences). There is no problem of the type that Feferman is referring to, when he brings up the matter of whether, for example, Paul Cohen is or is not a Platonist. The issues that might raise doubts in the minds of people like Cohen - or Gödel, or Feferman, or myself, for that matter - have to do with questions as to the absolute nature of the truth of mathematical assertions which refer to large infinite sets. Such sets may be nebulously defined or have some other questionable aspect in relation to them. It is not very important to any of the arguments that are given in *Shadows* whether very large infinite sets of this nature actually exist or whether they do not or whether or not it is a conventional matter whether they exist or not. Feferman seems to be suggesting that the type of Platonism that I claimed for Cohen (or Gödel) would require that for *no* such set could its existence be a conventional matter. I am certainly not claiming that - at least my own form of Platonism does not demand that I need necessarily go to such extremes. (Incidentally, I was speaking to someone recently, who knows Cohen, and he told me that he would certainly describe him as a Platonist. I am not sure where that, in itself, would leave us; but it is my direct personal impression that the considerable majority of working mathematicians are at least "weak" Platonists - which is quite enough. I should also refer Feferman to the informal survey of mathematicians reported on by Davis and Hersch in their book The Mathematical Experience, 1982, which confirms this impression.)

9.3 The issue as to the "existence" of some very large set might occasionally have a bearing on the truth or otherwise of certain P-sentences. Accordingly, a mathematician's belief with regard to such a P-sentence might be influenced by that mathematician's particular attitude to the existence of such a set. Questions of this nature were discussed in *Shadows*, Section 2.10, response to Q11, where it is concluded that there is no great issue to disturb significantly the Gödelian conclusion G. Feferman has not chosen to comment on this matter, so I suppose that he has no strong objection to my line ofreasoning.

# 10. What has Gödel's Theorem to do with Physics?

10.1 Maudlin questions the very basis of my contention that one can indeed deduce something important and new about the nature of physical laws from the actual behaviour of certain physical objects: human mathematicians.<6> As far as I can make out, his basic claim is that the computability, or otherwise, of mathematicians has no externally observable consequences. I find this claim to be a very strange one. He refers to what he calls the "Strong Argument", which he says is "clearly unsound". The Strong Argument contends that "no computer could reliably produce the visible outward motions of a conscious person" and, consequently, there must be something beyond computation in the behaviour of physical objects (e.g. humans). Maudlin's objection seems to rest on the *finiteness* of the total output of a human being. Whatever the total output of some human being might be (and his "human being" is "Penrose", of course!), that output would

indeed be finite. Therefore there would be some computer program which could, in principle at least, simulate that person's action.<7> This is a very odd line of reasoning, because it would invalidate any form of deduction about physical theory from observation whatsoever. The number of data points concerning observations of the solar system is finite, after all, so those data points could form the output of a sufficiently large computer, quite independently of any underlying physical theory. (Or they could be used to support a wrong theory with enough parameter freedom, such as the Ptolemaic theory, or even chariots in the sky.) I am tempted to reply to Maudlin by merely saying "be reasonable!"

10.2 Of course, canned answers could in principle provide any answer you want - even with infinite numbers of alternatives if the canning is allowed to be infinite. But the *whole point* of a Turing test (as Turing himself importantly understood) is that it takes the form of a question and answer session. It is simply not *practicable* to take into account all conceivable questions and follow-up questions and follow-up follow-up questions, etc. simply by storing all possible alternatives. (Anyone who has contemplated the task of writing a comprehensive CD-ROM program - or even a book such as *Shadows* in which one attempts to "second guess" all readers' possible counter-arguments - will begin to appreciate what I mean. There can be a significant complexity explosion even in the relatively short reasoning chains that are involved in such things.) Maudlin refers to this matter of complexity explosion, but he does not draw the appropriate conclusion from it.

10.3 My contention is that without any *genuine understanding* on the part of the computer, it will (at least in most cases) eventually be found out, when subjected to sensitive enough questioning. Trying to simulate intelligent responses by having mountains and mountains of stored-up information, using the programmer's best attempts to assimilate all possible alternatives, would be hopelessly inefficient. It appears that Maudlin believes that he has made a decisive logical ("in principle") point by bringing in the finiteness argument. But he is allowing his computer to have an exponentially larger finite storage limit than the finite limit that he imposes on the human (which is a general feature of the "canned response" approach), and this is totally unreasonable. Indeed, this "exponential" relationship involved in a canned response (or in what is called a "look-up table") *is* a decisive logical ("in principle") response to Maudlin's proposal. This applies both in the finite and in the (idealized) infinite case; for we have 2^alpha > alpha whether or not alpha is finite, and this inequality comes from the same kind of diagonal argument (Cantor's original argument) as that used in the Gödel theorem.

10.4 In fact the finiteness issue was discussed in *Shadows* (in the responses to Q7 and Q8 in Section 2.6), though from a slightly different angle. Maudlin does refer to this discussion, but he appears to misunderstand it. (Baars, in expressing his somewhat muddled parallel/serial worries about the "infinite memory" of a Turing machine is in effect, also addressing the "finiteness" issue, but he does not refer to my discussion of it, nor to the relevant Section 1.5.) In that discussion, I addressed the problem of how one might provide answers to mathematical questions - of, say, deciding the truth of P-sentences - by simply listing all the correct answers. In my response to Q7, I pointed out

that the very process of listing the answers required some means of forming reliable truth judgements. This matter has simply been ignored by Maudlin, yet it contains the whole point of the non-computability argument. In order to be able to list the *correct* answers to the P-sentences in his canned responses, Maudlin's *computer programmer* will need to possess the (non-computable) quality of understanding in order to provide what are actually the correct answers! When I said, in *Shadows* that "the odd against this are absurdly enormous", I was referring to the chances against providing the answers to mathematical problems of this nature without any understanding on the part of the programmer. Maudlin's situation is completely different, where he in effect presupposes that the programmer is allowed to have this understanding, and this completely begs the non-computability question.

10.5 There is, however, a somewhat related issue that has also been raised with me by other people: how could one actually *tell*, by observational means alone, whether or not the physical world behaves non-computably? (Here, I am leaving aside the question of the behaviour of extremely highly sophisticated physical objects like human beings; I am concerned with direct physical experiments and the like.) It seems to me that this issue is quite comparable to a somewhat related one, namely that of *determinism*. How could one tell by direct physical experiment whether or not the physical world is deterministic? Of course, one cannot tell - not just like that. Yet there is the common assertion that the classical behaviour of physical objects is indeed deterministic. What this means is that Newtonian *theory* (or Maxwell's *theory* or Einstein's *theory*) is deterministic; that can be shown mathematically. What one does is to design sophisticated experiments or observations to test the theory in other respects, and if the expectations of the theory are borne out, we conclude that various other things about that theory, such as the fact that it is indeed deterministic, ought also to hold for the behaviour of the universe (to the appropriate degree of approximation as is implied by the limits within which the theory has been shown to be valid). And so it will be with the new theory of physics that unites the classical and quantum levels and which, I maintain, will turn out to be a non-computable theory. Of course, I am at a disadvantage here, since this theory has yet to be discovered! But the general point is the same.

# 11. How Could Physics Actually Help?

11.1 Several commentators (Baars, Chalmers, Feferman, Maudlin) question the competence of *any* physical theory ever having anything of importance to say about mind, consciousness, qualia, etc. and Klein asks for clarification on this issue. According to Feferman, for example, my attempts to push the consciousness discussion in the direction of physics would merely be to replace one "nothing but" theory with another, i.e. to replace "the conscious mind is nothing but a computer" with "the conscious mind is nothing but sub-atomic physics". Other commentators, in effect, express similar worries. In fact, to describe things in the aforementioned way is rather to miss the point of what I am trying to say. I certainly do not expect to find any answers in sub-atomic physics, for example. What I am arguing for is a *radical upheaval* in the very basis of physical theory.

11.2 In most respects, this upheaval would have to have no observable effects, however. This might seem odd, but we have an important precedent. Einstein's general relativity, as regards most (indeed, almost all) of its observational consequences, is identical with Newton's theory of gravity. Yet, it indeed provided a radical upheaval in the very basis of physical theory. The concept of gravitational force is gone. the concept of a flat background Euclidean space is gone. The very fabric of space-time is warped, and the density of energy and momentum, in whatever form, directly influences the measure of this warping. The precise way in which the warping occurs describes gravity and tells us how matter is to move under its influence. Self-propagating ripples in this space-time fabric can occur, and carry away energy in a mysterious non-local way. Although for many years observational support for Einstein's theory was rather marginal, it can now be said that, in a clear-cut sense, Einstein's theory is confirmed to a precision of one part in one hundred million million - better than any other physical theory (see *Shadows*, Section 4.5).

11.3 What I am asking for is a revolution of (at least) similar proportions. It should represent as much of a change in our present-day ways of looking at quantum theory as general relativity represents a change from Newtonian theory. Some will argue, however, that even the profound changes that I have described above, which overturn the very basis of Newtonian physics, will do nothing to help us come to terms with the puzzle of mentality within such a physically determined universe. I do not deny the significance of that argument. But we do not yet know the very form that this new theory must take. It might have a character so different from that which we have become accustomed to in physical theory that mentality itself may not seem so remote from its form and structure. Moreover, quite apart from any considerations of mentality, there are, in my opinion, very powerful reasons coming from within physics itself for believing that such a revolution is necessary. (Baars, in particular, fails to appreciate this point when he says "there is yet nothing to revolt against".)

11.4 Einstein's theory was to do with the issue of how to describe the phenomenon of gravity - in its action in guiding the planets and the stars and the galaxies, and in the shaping of the large-scale structure of the universe. These phenomena do not directly relate to the processes which control the behaviour of our brains and which presumably actually underlie our mentality. What I am now asking for is a revolution that would operate at the very scales relevant to mental processes. Yet I am also arguing that the physical revolution we seek should actually be dependent upon the particular revolutionary changes that Einstein's theory already represented from the older Newtonian ideas about the nature of reality.

11.5 I know that this puzzles many people; in fact, it puzzles many *physicists* that I should seriously attempt to claim such a thing. For the scales at which gravitational interactions reign seem totally different from those which operate in the brain. A few words of explanation may well be helpful at this juncture. I am certainly not asking that

gravitational interactions (or "forces") should have any significance for the physical processes that are going on in the brain. The point is quite a different one. I am referring, instead, to the influences that Einstein's viewpoint with regard to gravity will have upon the very structure of quantum theory. Instead of quantum superpositions persisting for all time - as standard quantum theory would have us believe - such superpositions constitute a state which is *unstable* (see Penrose 1996). Moreover, this decay time can be computed, at least in certain very clear-cut situations. Yet, many physicists might well take the view that the time-scales, distance-scales, mass-scales, and energy-scales that would arise in any framework that purports to embody the union of Einstein's general relativity with quantum theory must be hopelessly wrong. Indeed the relevant time-scale (~10^-43 seconds) is some twenty orders of magnitude shorter that the briefest processes that are considered to take place in particle physics; the relevant space-scale (~10^-13 cm) is some twenty orders of magnitude smaller than the diameter of a proton; the relevant mass-scale (~10^-5 grams) is about the mass of a flea, which seems much too big; and the relevant energy scale (~10^18 ergs) is about what would be released in the explosion of a can of petrol. However, when one comes to examine the details, these figures conspire together (some being individually too small but others correspondingly too big) to produce an effect that is indeed of an eminently appropriate magnitude. (For details, see a forthcoming paper by Stuart Hameroff and myself (Hameroff and Penrose 1996).)

11.6 Again, many would argue that we shall still have come no closer to an understanding of mentality in physical terms. Perhaps, indeed, we will not have come a great deal closer. But I believe that some progress will have been made in an appropriate direction. The picture of quantum state reduction that this viewpoint is concerned with ("OR": objective state-reduction) involves the bifurcation and then selection of one out of several choices for the very shape of space-time. Moreover, there are fundamental issues arising here as to the nature of time and the apparent flow of time (see Section 13 below, in relation to Klein's commentary). I am not arguing that these issues will, in themselves, resolve the puzzles of human mentality. But I do claim that they could well point us in new directions of relevance to them, and this could change the very nature of the questions that the problems of mentality raise.

11.7 I think that people in AI, and perhaps a good many philosophers also, have a tendency to underestimate the importance of the *specific* nature of the physical laws that actually govern the behaviour of our universe. What reason do we really have to assume that mentality does not need these particular laws? Could consciousness arise in a world controlled by some arbitrarily chosen set of rules? Could it arise within scope of John Conway's "game of life" (Gardner 1970, Poundstone 1985), for example, as Moravec (1988) has suggested? Although the Conway rules for a "toy universe" are ingenious, they do not have the subtle sophistication of Newtonian mechanics - whose sophistication people often take for granted. Yet despite the extraordinary fruitfulness of Newtonian ideas, even they cannot explain something so basic as the nature and stability of atoms. We need quantum theory for that. And even quantum theory does not fully account for the behaviour of atoms, because its explanations require that curious hybrid of procedures of unitary (Schroedinger) evolution and quantum state-vector reduction (denoted in

*Shadows* by U and R, respectively) - procedures which are not really consistent with one another, I claim. Eventually, in order to explain even the stability and the specific nature of atoms, we shall need a better theory of physics than we have today, at the *fundamental* level.

11.8 There is no doubt that physics - and often the very detailed nature of the specific underlying physical laws - is essential to most of the sophisticated behaviour of the world we know. So why should the most sophisticated behaviour that we know of in the world, namely that of conscious living human beings, not also depend on the very detailed nature of those laws? As I have indicated above, we do not yet know the full nature of these laws, even in some of their most basic respects. A new theory is needed quite independently of any necessity for new laws to describe a universe that can support consciousness. However, physicists themselves often get carried away into thinking that they know everything that is needed - in principle, at least - for the behaviour of all things of relevance. There is a curious irony, here, in McDermott's quoting from *Shadows* p.373 "It is only the arrogance of the present age that leads so many to believe that we now know all the basic principles that can underlie all the subtleties of biological action." For he takes that remark to be aimed primarily at the AI community. In fact, the people I had primarily in mind were the (theoretical) *physicists*. I do not blame the biologists, or even AI researchers, when they take from the physicists a picture of the world commonly claimed to be almost final - bar some technical details that are irrelevant for the behaviour of macroscopic objects. But perhaps McDermott is right; some AI researchers seem to be nearly as arrogant as high-energy physicists (and with far less reason) - especially those AI researchers who claim that the deepest mystery of the physical world can be answered without any reference to the actual laws that govern that world!

11.9 I should make it clear, however, that I am certainly making no claim that the mystery of mentality can be resolved *merely* by finding the correct physical theory. I am sure that there are vital insights to be gained from psychology as well as from neuro-physiology and other aspects of biology. Baars seems to think that I am denying the existence of the *unconscious*, because there is no significant mention of it in *Shadows* (though there was some small reference to the unconscious mind in *The Emperor's New Mind*). I should like to reassure Baars that I fully accept both the existence of the unconscious and its importance to human behaviour. The only reason that the unconscious was not discussed in *Shadows* was that I had no contribution to make on the subject. I was concerned with the issue of *consciousness* directly, in particular in relation to the quality of understanding. However, I certainly agree that a complete picture cannot be obtained without the proper role of unconscious mentality being appreciated also.

# 12. State-Vector Reduction

12.1 Some commentators express worries in connection with my quantum state-vector proposals - whereby the quantum procedure R is to be replaced by some form of *objective* reduction, which I denote by OR. There are many misunderstandings here. Baars seems to think that I am taking the view that R has something to do with "observer

paradoxes", which is explicitly not my view, as I thought I had made clear in *Shadows*, Chapter 6. Klein does not make this mistake, but seems to take the view that the measurement problem (R) has (or ought to have) something directly to do with metaphysics. This is certainly different from my own "objective" standpoint with regard to R.

12.2 Maudlin complains that my "objections to Bohm's theory" (a theory that, in a sense, incorporates R) "are impossible to decipher from the text" - which is not surprising since I did not give them there - and that my "objections to the GRW theory are clearly not decisive". My objections to GRW (the OR scheme of Ghirardi, Rimini, and Weber, 1986) were not meant to be decisive. In my opinion, this scheme is a very interesting one, but it suffers from being somewhat ad hoc. What one needs (and I am sure that the authors of this scheme would not disagree) is some way of fitting the scheme in more convincingly with the rest of physics. In fact Diosi made a proposal in 1989 that could be regarded as a GRW-type model in which the ad hoc nature of the GRW parameters was removed by fixing them to be provided by the *quantum gravity* quantities referred to in Section 11, above. Diosi's model encountered difficulties, as was pointed out by Ghirardi, Grassi, and Rimini (1990), who also suggested a remedy, but at the expense of re-introducing another parameter. It should be said that in fact the Diosi-Ghirardi-Grassi-Rimini proposal is extremely close to the OR scheme that I was proposing in *Shadows* (and in Penrose 1995). These other authors do not mention non-computability (their proposal being entirely stochastic), but there is no essential incompatibility between both sets of ideas. In Sections 7.8 and 7.10 of *Shadows*, I give some reasons (admittedly far from conclusive) for anticipating that a full quantum gravity scheme of this nature might indeed be non-computable.

12.3 In his final paragraph, Maudlin seems to be complaining that the tentative OR proposals that are being promoted in *Shadows* do not solve all the problems of uniting quantum theory with relativity, and of explaining the problems of human cognition. He's not asking for much! These proposals were hardly intended to provide a complete theory (as anyone reading Section 7.12 of *Shadows* would surely appreciate) but merely to give some idea of the orders of magnitude involved in the collapse rate for such an OR theory - if such a theory *were* to be found.

12.4 Klein refers to the excellent little book QED of Feynman (1985) which introduces the basic rules of quantum theory (and quantum fieldtheory) with the minimum of fuss. However, Feynman never attempts to address the measurement problem in this book - which amounts to the issue of why (and when) do the quantum-level complex-valued amplitudes become classical-level real-valued probabilities, in the process of having there moduli squared. It might be worth mentioning that I read QED for ideas, just before embarking on writing my chapter on quantum mechanics in *The Emperor's New Mind*. However, I found that Feynman's approach was not altogether suitable for me because I needed to address the measurement problem in some detail, which Feynman avoided completely. Feynman has certainly worried about this problem, but he preferred not to

emphasize it in his writings. There is a historical point of interest, here. For it was actually Feynman's early worrying about the nature of the union of Einstein's general relativity with quantum mechanics (expressed in Feynman's contribution to the conference held in Chapel Hill in the 1950s, cf. Feynman et al 1995, Section 1.4, p. 15) that originally motivated Karolyhazy (1966) to seek an explanation of state-vector reduction in terms of gravitational effects (and Feynman also influenced me in the same way). Diosi's particular approach arose from the work of Karolyhazy's Budapest school.

12.5 Questions to do with "the overlap of states" referred to by Klein do not really resolve the measurement issue, and von Neumann's point about the difficulty of locating exactly where (or when) R takes place just emphasizes the subtlety of the R phenomenon. However Klein is completely right in pointing to the biological difficulties involved in maintaining quantum coherence within microtubules and, more seriously, in allowing this coherence to "leap the synaptic barrier". To see how this might be achieved is a fundamental problem for the type of scheme that I (in conjunction with Stuart Hameroff) have been proposing. Clearly more understanding is needed. (See Section 14 below, for a tentative suggestion in relation to this.)

12.6 I should point out a misunderstanding on the part of Maudlin. He seems to think that my "collapse theory offers a stochastic collapse postulate" and that it is concerned with "the exact timing of the collapses". I have nowhere said either of these things. I get the impression that Maudlin has been confused by the comparisons that I have been making between the suggested OR process and the phenomenon of the decay of an unstable particle (or unstable nucleus). But at no stage did I suggest that it would be in the precise *timing* of the decay of the quantum superposition that significant non-computability would occur. (But on rereading the relevant parts of *Shadows*, I realize that I was not at all specific there as to what I *did* mean.) Of course, since the detailed theory is not known, it is possible that there would be relevant non-computability in the timing also, but what I had in mind was something quite different, and certainly more relevant. The idea is that when collections of tubulin conformational movements become involved in coherent quantum superposition, there will come a point when the mass movements within the tubulin molecules are sufficient for OR to come into effect (without significant environmental disturbance). When that happens Nature must make a choice between the various collections of conformational states under superposition. It is not so much a question of *when*, but of *which* among the collection of superposed states Nature indeed chooses. Choices of this kind could actually influence the behaviour of a synapse. (There are various possibilities for this; for example, the particular collections of conformational states of tubulins in a microtubule might influence a dendritic spine, via the actin within the spine. Moreover, a great number of microtubules would be expected to act in concert - since a single OR state-selection process would act within many microtubules all at once. However, there is no point in trying to be too specific at this stage.) It would be in the *particular* choice that Nature makes that the non-computability could enter significantly, and this particular choice (global over a significantly large area of the brain, probably involving at least thousands of neurons) could result in subtle collective changes of synapse strengths all at once. (See Hameroff and Penrose 1996.)

# 13. Free Will

13.1 What kind of a theory might it be that determines these choices? Many people who are unhappy with computationalism would be just as unhappy with any other type of mathematical scheme for determining them. For they might argue that it is here that "free will" makes its entry, and they would be unhappy that their free-will choices could be determined by *any* kind of mathematics. My own view would be to wait and see what kind of non-computable scheme ultimately emerges. Perhaps a sophisticated enough mathematical scheme will turn out not to be so incompatible with our (feelings of) free will. However, McCarthy takes the view that I am "quite confused" about free will, and that my ideas are "not repairable". I am not really clear about which of my confused ideas McCarthy is referring to. In *Shadows*, I did not say much about the issue of free will, except to raise certain issues. Indeed, I am not at all sure what my views on the subject actually are. Perhaps that means that I *am* confused, but I do not see that these ideas are remotely well enough defined to be irreparable!

13.2 As I remarked above, most people would probably take the view that if there is *any* kind of mathematical theory precisely determining the way we behave, then there is no free will. But, as I have indicated, I am not so sure about this. The answer could depend on the very nature of this mathematical theory. The theory would certainly have to be non-computable (according to my own considerations), but much more than this. We recall from the discussion given in Section 3 above (McCullough, Chalmers, and Section 7.9 of *Shadows*) that the Gödel diagonalization procedure can be applied to systems much more general than merely computational ones. Thus, my arguments would equally imply that our missing theory must be not just non-computational, but also beyond (or at least different from) Turing's notion of oracle computation. (An oracle computation is what is achieved by a Turing machine to which an additional command is appended: "if the pth Turing machine acting on the number q eventually halts, do A; if it doesn't, do B".) Again, we can consider second-order oracle machines (which can assess whether first-order oracle machines ever halt), and the diagonalization still applies. So the missing theory is not a second-order oracle theory either. The same applies to even higher-order oracles. Indeed the missing theory cannot be an oracle theory of order alpha for any computable ordinal number alpha. As far as I can make out, this is not the limit of it either. Diagonalization can be applied in very general circumstances indeed. We enter very nebulous areas of mathematical logic. It seems that the quality of "understanding" - which is what this discussion is effectively all about - is something very mysterious. Consequently, any theory of the physical world which is capable of accommodating beings that are capable of genuine understanding must itself be in a position to cope with such subtleties.

13.3 As a side comment, I should remark that this form of "repeated Gödelization" is somewhat related to, but not at all the same as, that referred to by McCarthy and McCullough, who both describe the process whereby sound extensions of a sound formal system can be obtained, corresponding to any computable ordinal alpha. This procedure

was described in *Shadows* Section 2.10, answer to query Q19. I am not quite sure why they go to the trouble to repeat this argument, with no reference to my own discussion. The conclusion, noted in *Shadows*, that "repeated Gödelization does not provide us with a mechanical procedure for establishing the truth of P-sentences" is confirmed by Feferman. (As far as I can make out, Feferman's comments about how his work extended that of Turing's are related to the considerations of the previous paragraph, above.)

13.4 The issue of free will is related also to the experiments of Libet (1992), (and to earlier experiments of Deeke, Groetzinger, and Kornhuber (1976) and also Grey Walter) that are referred to by Klein. These experiments suggest a delay of the order of a second, in an entirely volitional act, between the first indications of mental activity (as evidenced by brainwave studies) and the final willed (say) finger movement. Klein calls into question my puzzlements, expressed in *Shadows* Section 7.11, concerning the seeming slowness of consciousness - but as far as I can make out, he has misunderstood my point (as, I believe, did Ian Glynn before him, in the 1990 article that Klein refers to). Klein says that there are "no surprises" in the fact that there is "substantial unconscious processing" (in fact, of the order of a second's worth) before "the subjective awareness of the decision to act" takes place, and that only about one fifth of a second's delay occurs before the motor act. But it was *precisely* the length of time involved in the unconscious processing that was worrying me. (Or at least this was *one* of the two things that was worrying me; the other had to do with the *passive* half-second delay that Klein also regards as an over-estimate - to which I shall return below.)

13.5 If consciousness has any active role to play in a response to an external stimulus, then it is no good for the unconscious to have already lined up the action a second ahead of time unless the unconscious already "knew" what decision that the conscious mind was going to take. Klein asserts that I am referring to a "stimulus-response situation", for which the response could be much more rapid and essentially entirely *unconscious*. In fact I am *not* considering situations of this kind, but those in which it is necessary for consciousness to come into play in order that it can actively influence the outcome. If the "free will" of the conscious mind is allowed to come into play, then surely it would be necessary for *all* processes that need to be involved, whether they be the conscious ones or the preparatory unconscious activities, to take place *after* the external stimulus occurs. This leads us to a about a second's delay for a consciously influenced response.

13.6 The extra half second comes from Libet's other (passive) experiment (Libet et al 1979), which Klein argues may be too long for stimuli that are significantly greater than threshold. I do not wish to argue this point, since I am not aware of the relevant figures (the 100 msec figure referred to by Klein being not relevant to the situation I was considering, as far as I can see). A one-second's delay for a consciously controlled free-willed response seems already an inordinately long time.

13.7 I am not claiming that these considerations are decisive, in any way, as indications

that the quantum/relativity puzzles concerning the nature of time and causality have significance for our consciousness and perception of "the flow of time". However, it seems to me that it is quite possible that there is something very odd going on concerning the timing of conscious events, if only for the reasons indicated in *Shadows* Section 7.11 that the role of time with respect to consciousness is quite different from its role in physics - in that it is only with the phenomenon of consciousness that time seems to "flow". I certainly hope that more experiments of the types that Libet and his associates have been performing will be carried out in the future. I suspect that there may be further surprises in store.

# 14. Some Remarks on Biology

14.1 Many people have expressed reservations (of widely differing degrees) concerning the biological speculations put forward in *Shadows*. I have referred (in Section 12 above) to Klein's worries about the difficulty of maintaining quantum coherence within individual microtubules and, moreover, of this quantum coherence straddling, simultaneously, a great number of microtubules within collections of separated neurons. I agree with Klein that it would be an extraordinary challenge to see how such organization might be achieved. Yet, I maintain that somehow Nature must indeed have accomplished this extremely remarkable task. In this section, I shall try to address this issue further, and also address some of the other objections that have come my way. I shall also relate several new things, in relation to these issues, that I have learned since I wrote *Shadows*.

14.2 One complaint that I have heard is that the biological purpose of microtubules within cells is "already known", namely that they are there to provide the "tracks" along which molecules (generally known as "organelles") are transported from one part of a cell to another; and they grow, shrink, or bend in ways that are designed to influence the movement of cells. Moreover, so these arguments might continue, their tubelike construction is to give them structural strength - so there is no need to ask for a separate purpose for the tubes, such as to isolate some kind of quantum-coherent activity taking place within the tubes, from the outside environment. I do not doubt that microtubules indeed perform the tasks that they are presently believed to perform, and many more besides. But that is no argument against their *also* serving the additional purposes that I require of them. We know of many instances where Nature uses the same structure for many different purposes. We know that mammalian noses, for example, filter substances from the air before it reaches the lungs (not to mention their importance to the sense of smell). Yet this is no argument against elephants *also* using their noses in delicate ways to pick up objects from the ground!

14.3 A more serious argument is the lack of direct evidence for the type of "cellular automaton" activity that Hameroff and his colleagues have been arguing for in patterns of tubulin conformations along microtubules. The existence of some kind of activity of this general nature is indeed part of the general picture that Hameroff and I would require for our model of the physical processes underlying consciousness. To obtain direct

experimental support for this kind of activity would be a key issue, and I certainly hope that it will be possible to design experiments to test it. Experimental support for the existence of some kind of quantum coherence within microtubules is a matter of even greater importance for the ideas that I have been promoting in *Shadows*. There is no doubt that definitive experiments would be difficult to perform, especially since there is a distinct possibility that the relevant effects might require microtubules *in vivo* rather than *in vitro*. I have been informed by Guenther Albrecht-Buehler that there is some kind of coating (analogous to the myelin sheaths of neurons) that microtubules have *in vivo* which tends not to be present *in vitro*.

14.4 On the theoretical side, some progress has been made. Work by Tuszynski et al (1996) gives theoretical support for information processing (of the Hameroff type) to be possible, within an appropriate temperature range, provided that the microtubule possesses the structure of what is known as the "A-lattice", which is indeed the structure depicted in *Shadows* figures 7.4, 7.8, 7.9, on pp. 359, 363. However, the work of Mandelkow and Mandelkow (1994) indicates that many (perhaps most) microtubules seem to have a somewhat different structure, known as the "B-lattice", in which there is a "seam" running the length of the microtubule. Tuszynski et al argue that the B-lattice is not capable of sustaining Hameroff-type information processing, but it may well be appropriate for the transporting of organelles. It would be extremely interesting to have information about which kind of lattice structure is prevalent in axons, dendrites, non-neuronal cells, etc.

14.5 With regard to the theoretical possibility of quantum coherence within microtubules, the model of Jibu et al (1994) seems well-founded, in which super-radiance effects are anticipated within microtubules (analogous to the activity of a laser), where the electromagnetic field interacts with ordered water. For this process to occur, it would be necessary for the water within the tubes actually to adopt this ordered structure, and to be appropriately free of the wrong kind of impurities, such as chloride ions. (Apparently, sodium, calcium, magnesium, and potassium ions, in low enough concentrations should not disturb the ordering.) It should be mentioned, however, that the type of coherent activity that is anticipated in the model of Jibu et al may not be sufficient for my purposes. Though it is a necessarily quantum effect, it is not, as it stands, a *quantum-coherent* effect of the type that my arguments require. Genuine quantum coherence seems to be necessary in order that the quantum/classical borderline can be probed, where the (non-computable) effect of the missing OR theory can significantly make its mark. (This comment has relevance to Klein's query at the end of his Section 1. *Classical* coherence in the brain may well occur, but it does not provide an opening for non-computational activity, which I argue is a characteristic feature of consciousness.) The Jibu et al mechanism may be part (though not all) of what is needed.

14.6 An interesting possibility has come my way, which may conceivably have relevance to the question of how quantum coherence might get conveyed between one neuron and another (a question raised by Klein). As noted in *Shadows*, Figs. 7.11, 7.12 on pp. 365,

366, there are some particular molecules (clathrins) that inhabit synaptic boutons, which have the highly symmetrical structure of a *truncated icosahedron* (like a modern soccer ball). These clathrin molecules have importance in the release of neurotransmitter chemicals at synapses (whereby the nerve signals are transmitted from neuron to neuron). Although I do not have specific suggestions to make here, I am struck by the extraordinary symmetry of these molecules. It has been brought to my attention (by Roy Douglas, cf. Douglas and Rutherford 1995) that, according to the Jahn-Teller effect, such highly symmetrical molecules would have a large energy gap between the lowest quantum energy level and the next. This lowest level would be highly degenerate, and there would be interesting quantum-mechanical effects when this degeneracy is broken.

14.7 Energy gaps and symmetry breaking, of this general nature, are central to the understanding of superconductivity - and superconductivity is one of the few clear phenomena in which large-scale quantum coherence takes place. Known observationally since 1911, and explained quantum-mechanically in 1957, superconductivity had been thought originally to be an exclusively very low-temperature phenomenon, occurring only at a few degrees above absolute zero. It is now known to occur at much higher temperatures of -158 degrees Celsius, or perhaps even -23 degrees (although this is not properly explained). It does not seem to be out of the question that there might be similar effects at the somewhat higher temperatures of microtubules. Perhaps there are understandings to be obtained about the behaviour of microtubules from the experimental insights gained from such high-temperature superconductors.

14.8 Another question frequently asked is: what's so special about neuronal microtubules, as opposed to those, say, in liver cells? In other words, why isn't your liver conscious? In answer to this, it should be said that the organization of microtubules in neurons is quite different from that in other cells. In most cells, microtubules are organized radially, from a central region (close to the nucleus) called the *centrosome*. In neurons, this is not the case, and they lie essentially parallel with one another along the axons and dendrites. The total mass of microtubules within neurons seems to be much greater than in other cells, and they are mainly stable structures, rather than in most cells, where they continually polymerize and depolymerize (grow and shrink). Of course, there is much to be learned about the respective roles of microtubules in neurons and in other cells, but there does seem to be clear enough evidence for an essentially distinct role for (some of) those in neurons. (The A-lattice/B-lattice question would seem to be of importance here also.)

14.9 In this connection, I should mention something of considerable interest and relevance that I learned recently from Guenther Albrecht-Buehler (1981, 1991), which concerns the role of the *centriole*, that curious "T" structure (roughly illustrated in *Shadows*, Fig. 7.5, on p.360), consisting of two cylinders resembling rolled-up venetian blinds, constructed from microtubules and other connectingsubstances, which lies within the centrosome. In *Shadows*, I had adopted the common view that the centrosome acts in some way as the "control centre" of the cytoskeleton of an ordinary cell (not a neuron), and that it initiates cell division. However Albrecht-Buehler's idea about the role of the

centriole is very different. He argues, convincingly, in my opinion, that the centriole is the *eye* of the cell, and that it is sensitive to infra-red light with very good directional capabilities. (Two angular coordinates are needed for identifying the direction of a source. Each of the two cylinders provides one angular coordinate.) Impressive videos of fibroblast cells provide a convincing demonstration of the ability of these cells to pinpoint the direction of an infra-red light source. This also provides some remarkable evidence for individual cells having considerable information-processing abilities, which is at variance with current dogma. One may well ask where the "brain" of a single cell might be located. Perhaps its structure of microtubules can serve such a purpose, but it does seem that the centrosome itself must have some central organizing role. In a single (non-neuronal) cell, the microtubules emanate from the centrosome. I gather from Albrecht-Buehler that the specific contents of the centrosome are not known. It seems that it would be important to know what indeed is going on in the centrosome. Does it have some information-processing capabilities? Is there conceivably some structure there that is capable of sustaining quantum coherence in any form? The answers to questions of this nature could have considerable importance.

14.10 I should make clear that I am not arguing for any consciousness (or consciousness of any significant degree) to be present for individual cells. But according to the views that I have been putting forward, some of the ingredients that are needed for actual consciousness ought already to be present at the cellular level. Individual cells can behave in strikingly sophisticated ways, and I find it very hard to see how their behaviour can be explained along entirely conventional (classical) lines.

14.11 All this notwithstanding, there is the question of whether microtubules are indeed necessary for consciousness to be present in human beings or other animals. An argument that I have heard presented - as though it were a conclusive refutation of this contention (cf. Grush and Churchland 1995, Edelman 1995) - is that the drug colchicine, which is given as a treatment for gout, depolymerizes microtubules, yet it does not influence the mental state; moreover, when colchicine is delivered directly to the brains of experimental animals, they appear to remain conscious. This can be answered (cf. Penrose and Hameroff 1995 and references contained therein) by pointing out that: (1) the blood-brain barrier is not significantly breached in the gout patients, so their neuronal microtubules are not disturbed, and (2) in any case, most brain microtubules, unlike those in non-neuronal cells, are stable structures that do not undergo cycles of polymerization and depolymerization, and are *resistant* to colchicine. A minority of neuronal microtubules - those involved in restructuring synaptic connections - are involved in such activity, however, and these could be affected by colchicine. Indeed, the experimental animals referred to above suffer a kind of "dementia", similar to Alzheimer's disease (Bensimon and Chernat 1991), a disease which has itself been linked to microtubule disruption (e.g. Matsuyama and Jarvik 1992).

14.12 Of course, there is the additional issue of how we could know whether a demented rat is or is not conscious. We must return to the question of what consciousness is and what are its external manifestations.

# 15. What is Consciousness?

15.1 In *Shadows* (cf. Section 1.12) I concentrated specifically on the quality of "understanding" as a particular manifestation of consciousness - a quality that would make its characteristic mark on external behaviour as well as being an internal manifestation of mentality. Only with respect to the quality of understanding have I been able to argue for non-computable ingredients being necessary. But, in my view, non-computable physical processes must also be essential for other aspects of conscious mentality.

15.2 Consciousness has its active aspects - basically the "free-will" issues that were considered in Section 13 above - and it has its passive aspects, which have to do with awareness and the vexed issue of *qualia*. Understanding fits somewhere between the two. In my view, anything that sheds some light on the problem of how a physical system can exhibit understanding must inevitably also shed some light on the "free-will" and "qualia" problems. Moreover, the issue of "understanding" seems to me to be one of the more tangible aspects of consciousness. I do not see how to say much that is scientifically useful about the qualities of "free will" or "awareness", but "understanding" is something that we can work with. Klein raises the issue of Wilczek's challenge of "looking for perceptual feats that humans can do more efficiently than robots". My answer would be: anything in which the quality of *understanding* is important. A good example is the chess problem presented in *Shadows* Fig. 1.7 on p.46. This problem was one of a series composed by William Hartston and David Norwood, consisting of chess problems, some of which were designed to be easy for humans but hard for computers (best solved using "understanding") and others, the other way about (best solved by "trying all possibilities"). This "Turing test" showed a virtually complete separation between humans and computers.

15.3 As must be clear from the preceding remarks, I do not believe that any real progress will be achieved towards solving the mysteries of how mental phenomena fit in with the physical universe until there are some important changes in our picture of physical reality. Perhaps these developments will lead to a theory in which "consciousness" finds some place within the purely physical descriptions of the world. One is reminded of such ideas as "panpsychism" (like those of Leibniz, Spinoza, or Whitehead), where consciousness may play its part within the processes of physical action at its deepest levels. I do not have strong opinions as to the significance of such ideas, mainly because I have not studied them in detail. But I suspect the truth to have a much more compelling grandeur to it than any set of ideas that I have seen so far.

15.4 Thus, I certainly do not go along with the "no big deal" viewpoint of McDermott, a viewpoint which is, in effect shared by Baars when he takes the view that the solutions are all to be found within psycho-biology, without any significant change in our physical world-view being necessary. I think that Baars grossly underestimates the force of the arguments from logic. (And although, as a mathematician, I frequently make use of

"variables", I have no idea what Baars means by "treating consciousness as a variable".) I fully accept that there are invaluable insights to be gained from the areas of psychology and biology. But, important though these areas are, I was not so much concerned with them as with physics in *Shadows*. For I believe that it is *also* fundamentally important to see whether our present physical world-view is in fact adequate for accommodating, in any way, the phenomenon of consciousness. I have extensively put the case here, and in *Shadows*, that it is not. The arguments from logic and from physics are not counter to those from psychology and biology, but complementary to them.

15.5 Likewise Moravec and McCarthy appear to belong to the "no big deal" school. McCarthy puts forward various suggestions for the circumstances under which he would consider that "consciousness" occurs. These are all within the computational model, so it is clear from this that I am not in agreement with him that his computer systems, acting according to his criteria, are *actually* conscious (in the sense that one could actually *be* such a system). Again, I fear that McCarthy does not appreciate the force of the logical arguments that I have given, which inform us that the quality of "understanding" cannot be accommodated within the computational model. It is easy to suggest *definitions* within the computational model (as McCarthy does) of such things as "consciousness", "awareness", "self-awareness", "intentions", "beliefs", "understanding", and "free will". But such definitions need not convey to us (and do not convey to me) any conviction that the corresponding mental qualities that humans actually possess are in any real sense captured by computational definitions of this nature. As I have argued extensively above, the actual quality of human *understanding* cannot be captured within *any* purely computational scheme. So it is clear that I cannot be in agreement with all of McCarthy's definitions.

15.6 Chalmers raises the issue of the distinction between "simulating", "evoking", and "explaining". I agree with him that there are indeed distinctions to be made. But I feel that these are distinctions that have more importance to a philosopher than to a scientist. Whilst I did myself distinguish between the possibility of "simulating" and of "evoking" consciousness in my "viewpoint B" and "viewpoint A" distinction in *Shadows* Section 1.3, I think that the normal scientific (as opposed to philosophical) stance would be to concentrate on what can be externally observed, so a system which succeeds in simulating the outward effects of consciousness would be suspected also of evoking them. Thus, my "viewpoint B" might not be a happy one for a hard-nosed scientist.

15.7 Likewise, a modern scientist might have trouble producing an "explanation" for something other than by producing a theory that in principle provides a (mathematical) "simulation" of that thing. It seems to me, therefore, that the "scientific" viewpoints (in this sense) are those who hold to the same position, with regard to computationalism, in respect of all three of the simulate/evoke/explain issues. And since Chalmers's "N" represents a denial of the competence of physics with regard to conscious mentality, this implies that we would be left with just "CCC" (computationalism all down the line) or "PPP" (non-computational physics all down the line). Thus, when I wear my scientist's

hat, I am unable to understand why someone (such as Chalmers) can hold to different positions with regard to the simulate/evoke/explain issues, although when I wear my philosopher's hat, I can partly appreciate his point. However, I wear my scientist's hat much more frequently than my philosopher's hat!

15.8 But sometimes I try to wear both hats at once. The arguments in *Shadows* were concerned almost entirely with the "simulate" issue with regard to human mentality. I hope that those who study, with a genuinely open mind, the arguments given there (and the further discussions above) will come to accept that a non-computational physics will be needed in order even to simulate the actions of a conscious being. There are, in any case, powerful reasons for believing that profound changes in our physical world-view are in the offing. For the resulting science to be non-computable to the degree that seems to be required, we may well find the need for a science that is so different from the science of today that the *evoke* and *explain* issues with regard to mentality may finally find natural explanations.

# Acknowledgement

I am grateful to the National Science Foundation for support under contract PHY 93-96246.

# Notes

<1> As I understand it, there was to have been a tenth commentary, explicitly addressing a number of biological points, but unfortunately this article did not materialize. Nevertheless, in Section 14, I shall find it helpful to address a number of biological criticisms that have come to my attention.

<2> Most book-reviewers seem to have missed this argument too. In particular, Hilary Putnam, in his widely quoted review of *Shadows* in the Sunday New York Times Book Section (Putnam (1994) and reprinted, for some reason, in the Bulletin of the American Mathematical Society, Putnam (1995)) not only completely missed this argument, but tries to claim that I have not considered other issues that I have, in fact, discussed in great detail. The matters are thoroughly discussed again in Sections 6 and 7 of this reply.

<3> Dennett, in his 1995 book *Darwin's Dangerous Idea*, seems to be trying to make out that I do not believe that human abilities can have arisen by the process of natural selection, since I do not believe in computationalism. This is a very strange contention. Provided that the non-computational ingredients are present in nature, there is nothing against natural selection having made use of them - and it is my own contention that this is indeed how things were. If Dennett is arguing that I do not believe that natural selection provides the sole explanation of the origin of human mentality, then he is right. In particular, some specific laws of physics are also needed, within whose scope natural

selection must operate. But that, in itself, is hardly a radical position, nor an unscientific one!

<4> It appears that some people, on reading the section entitled "Contact with Plato's world" in Chapter 10 of *The Emperor's New Mind*, have picked up the curious view that I believe that mathematicians obtain their mathematical knowledge by use of some direct mystical quality not possessed by ordinary mortals (see Grush and Churchland 1995, for example), and even that I may be claiming for myself a particularly unique such quality! This is a complete misreading of what I had intended in that section; for I was simply trying to find some explanation of the fact that different mathematicians can communicate a mathematical truth from one to another even though their modes of thinking may be totally dissimilar. I was arguing merely that the mathematical truths that each mathematicians may be groping for are "external" to each of them - these truths being "inhabitants of Plato's timeless world". I was certainly not arguing for a fundamentally *particular* quality of "direct Platonic contact" to be possessed only by certain individuals. I was referring simply to the general qualities of "understanding" (or "insight") which are in principle available to all thinking individuals (though they may perhaps come somewhat more *easily* to come individuals than to others). These qualities are not mystical - but as Gödel's theorem shows, there is indeed something rather mysterious about them.

<5> See, for example, Grush and Churchland (1995), and for a reply, Penrose and Hameroff (1995).

<6> John Searle, in his interesting recent review of *Shadows* in the *New York Review of Books* (November 2, 1995), seems to be making a somewhat similar point. However, he does not appear to have grasped, properly, the key notion of *non-computability* (and the fact that it has observational manifestations). See, in particular, the discussions given here (Section 7, 10, 11, 15) and also Sections 3.2 onward of *Shadows*, which explicitly address the conscious/unconscious issue that Searle raises, with regard to an algorothmic basis for mathematical understanding. The aforementioned sections may serve to clarify my own position concerning what he claims are my "fallacies".

<7> F.J.Tipler, in a review of *The Emperor's New Mind*, used a similar "finiteness" argument, specifically referring to the (absurdly large) Bekenstein bound on the information that can be stored within an object of a given (say human) size. I explicitly addressed this argument from finiteness in my responses to Q7 and Q8 in Section 2.6 of *Shadows*. However, I had deliberately desisted from indicating the personal individuals whom I had in mind (in this case Tipler) as putting forward the various specific "queries" that I was responding to. There is a certain irony for me in this, because in his review of *Shadows* in *Physics World*, Tipler chastised me, claiming that I did "not even mention ... the Bekenstein bound" (Tipler 1994), not having even noticed this section of the book that had been specifically aimed at his own arguments!

# References


Albrecht-Buehler, G. (1981). Does the geometric design of centrioles imply their function? *Cell Motility, 1,* 237-245.

Albrecht-Buehler, G. (1991). Surface extensions of 3T3 cells towards distant infrared light sources. *Journal of Cell Biology, 114,* 493-502.

Bensimon, G. and Chernat, R. (1991). Microtubule disruption and cognitive defects: effects of colchicine on learning behavior in rats. *Pharmacology and Biochemistry of Behavior, 38,* 141-5.

Davis, P.J. and Hersch, R. (1982). *The Mathematical Experience*. Harvester Press.

Deeke, L., Groetzinger, B., and Kornhuber. H.H. (1976). Voluntary finger movements in man: Cerebral potentials and theory. *Biology and Cybernetics, 23,* 99.

Dennett, D.C. (1995). *Darwin's Dangerous Idea*. New York: Simon and Schuster.

Devlin, K. (1988). *Mathematics: The New Golden Age*. London: Penguin Books.

Diosi, L. (1989). Models for universal reduction of macroscopic quantum fluctuations. *Physical Review, A40,* 1165-74.

Douglas, R.R. and Rutherford, A.R. (1995). Pseudorotations in molecules I: Electronic triplets. To appear.

Edelman, G. (1995). Quoted on p. 323 of *Frontiers of Complexity*, by Peter Coveney and Roger Highfield. London: Faber and Faber.

Feynman, R.P. (1985). *QED: The Strange Theory of Light and Matter*. Princeton: Princeton University Press.

Feynman, R.P., Morinogo, F.B., and Wagner, W.G. (1995). *Feynman Lectures on Gravitation*. Reading, MA: Addison-Wesley.

Gardner, M. (1970). Mathematical games: the fantastic combinations of John Conway's new solitaire game 'Life'. *Scientific American, 223,* 120-123.

Ghirardi, G.C., Rimini, A., & Weber, T. (1986). Unified dynamics for microscopic and macroscopic systems. *Physical Review, D34,* 470.

Ghirardi, G.C., Grassi, R., & Rimini, A. (1990). Continuous-spontaneous-reduction model involving gravity. *Physical Review, A42,* 1057-64.

Grush, R. and Churchland, P.S. (1995). Gaps in Penrose's toilings. *Journal of Consciousness Studies, 2,* 10-29.

Hameroff, S.R. (1987). *Ultimate Computing: Biomolecular Consciousness and Nano-Technology*. Amsterdam: North Holland.

Hameroff, S.R. and Penrose, R. (1996). Orchestrated reduction of quantum coherence in brain microtubules - a model for consciousness. In S Hameroff, A. Kaszniak and A. Scott (Eds.) *Toward a Science of Consciousness*. Cambridge, MA: MIT Press.

Hameroff, S.R. and Watt, R.C. (1982). Information processing in microtubules. *Journal of Theoretical Biology, 98,* 549-61.

Jibu, M., Hagan, S., Hameroff, S.R., Pribram, K.H., Yasue, K. (1994). Quantum optical coherence in cytoskeletal microtubules: implications for brain function. *BioSystems, 32,* 195-209.

Karolyhazy, F. (1966). *Nuovo Cim., A42,* 390.

Karolyhazy, F. (1974). Gravitation and quantum mechanics of macroscopic bodies. *Magyar Fizikai Polyoirat, 12,* 24.

Libet, B. (1992). The neural time-factor in perception, volition and free will. *Review de Metaphysique et de Morale, 2,* 255-72.

Libet, B., Wright, E.W. Jr., Feinstein, B. and Pearl, D.K. (1979). Subjective referral of the timing for a conscious sensory experience. *Brain, 102,* 193-224.

Mandelkow, E-M. and Mandelkow, E. (1994) Microtubule structure. *Current Opinions Structural Biology, 4,* 171-179.

Matsuyama, S.S. and Jarvik. L.F. (1992). Hypothesis: Microtubules, a key to Alzheimer's disease. *Proceedings of the National Academy of Science USA, 86,* 8152-6.

McDermott, D. (1990). Computation and consciousness. *Behavioral and Brain Sciences, 13(4),* 676.

Moravec, H. (1988) *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, MA: Harvard University Press.

Penrose, R. (1989). *The Emperor's New Mind.* Oxford: Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind.* Oxford: Oxford University Press.

Penrose, R. (in press). On gravity's role in quantum state reduction. *General Relativity and Gravitation*.

Penrose, R. and Hameroff, S. (1995). What 'gaps'? - reply to Grush and Churchland. *Journal of Consciousness Studies, 2,* 99-112.

Poundstone, W. (1985). *The Recursive Universe: Cosmic Complexity and the Limits of Scientific Knowledge.* Oxford: Oxford University Press.

Putnam, H. (1994). The best of all possible brains? *New York Times Book Review, Nov. 20,* 7-8.

Putnam, H. (1995). Review of *Shadows of the Mind*, by Roger Penrose. *Bulletin of the American Mathematical Society, 32,* 370-373.

Tipler, F.J. (1994) Can a computer think? Part II. *Physics World, December,* 51-52

Tusznyski, J., Trpisova, B., Sept, D., and Sataric, M.V. (1996). Microtubular self-organization and information processing capabilities. In S. Hameroff, A. Kaszniak, and A. Scott (Eds.) *Toward a Science of Consciousness*. Cambridge, MA: MIT Press.

Found Phys (2014) 44:557–575
DOI 10.1007/s10701-013-9770-0

# On the Gravitization of Quantum Mechanics 1: Quantum State Reduction

**Roger Penrose**

Received: 7 September 2013 / Accepted: 3 December 2013 / Published online: 11 January 2014
© The Author(s) 2014. This article is published with open access at Springerlink.com

**Abstract** This paper argues that the case for "gravitizing" quantum theory is at least as strong as that for quantizing gravity. Accordingly, the principles of general relativity must influence, and actually *change*, the very formalism of quantum mechanics. Most particularly, an "Einsteinian", rather than a "Newtonian" treatment of the gravitational field should be adopted, in a quantum system, in order that the principle of equivalence be fully respected. This leads to an expectation that quantum superpositions of states involving a significant mass displacement should have a finite lifetime, in accordance with a proposal previously put forward by Diósi and the author.

**Keywords** Quantum theory · Linear superposition · Measurement · Gravitation · Principle of equivalence · Diósi-Penrose state reduction

The title of this article (and of its companion, [16]) contains the phrase "gravitization of quantum Mechanics" in contrast to the more usual "quantization of gravity". This reversal of wording is deliberate, of course, indicating my concern with the bringing of quantum theory more in line with the principles of Einstein's general relativity, rather than attempting to make Einstein's theory—or any more amenable theory of gravity—into line with those of quantum mechanics (or quantum field theory).

Why am I phrasing things this way around rather than attempting to move forward within that more familiar grand programme of "quantization of gravity"? I think that people tend to regard the great twentieth century revolution of *quantum theory*, as a more fundamental scheme of things than gravitational theory. Indeed, quantum mechanics, strange as its basic principles seem to be, has no evidence against it from accepted experiment or observation, and people tend to argue that this theory is so well

R. Penrose (✉)
Mathematical Institute, Oxford, UK
e-mail: penroad@wadh.ox.ac.uk

established now that one must try to bring the *whole* of physics within its compass. Yet, that other great twentieth century revolution, namely the *general theory of relativity*, is also a fundamental scheme of things which, strange as *its* basic principles seem to be, also has no confirmed experiments or observations that tell against it—provided that we comply with Einstein's introduction of a cosmological constant $\Lambda$, in 1917, which now appears to be needed to explain the very large-scale behavior of cosmology. So why give quantum theory pride of place in this proposed union?

It is true that there are far more phenomena, on a human scale, which are found to require quantum mechanics for their explanations, than are those that require general relativity for their explanations—far, far more. Also, I think it is felt that since quantum mechanics normally refers to very small things, and general relativity to very big things—and we tend to think of big things as being made up of small things—then the theory of the small things, i.e. quantum mechanics, must be the more fundamental. However I think that this is misleading. Moreover, there is another issue that I regard as far more important than this issue of *size*, namely that of the *consistency* of the theory.

We know of the problems of the infinities that so frequently tend to arise in the quantum theory of fields, and much work is geared to trying to eliminate or, at least, deal with these infinities in a consistent and appropriate way. But, of course, general relativity also has its infinities, these arising inevitably in the gravitational collapse to a black hole, or perhaps of an entire universe. These are serious issues, representing potential mathematical inconsistencies in the global applicability of each theory. These are issues that certainly do need attending to, and it is often argued that the infinities of one theory might be alleviated if the principles of the other theory can be appropriately brought to bear on them.

But quantum mechanics also has another basic inconsistency, which applies to the very basis of its founding principles. This is usually referred to as the "measurement problem", though I tend to prefer Tony Leggett's terminology: the "measurement *paradox*". We can think of this as a fundamental conflict between two of the very foundational principles of quantum mechanics, namely *linearity* and *measurement*. Let us have a look at these issues, starting with linearity, which is an essential part of the *unitary* evolution **U** of the quantum state-vector. In Fig. 1, we envisage a high-energy photon aimed at some brown object, which ejects various other objects as soon as the photon hits it. In Fig. 2, a mirror is inserted in the path of the photon, and the reflected photon hits a green object instead, which ejects a quite different family of objects when hit by the photon. In Fig. 3, the mirror is replaced by a beam-splitter (effectively a half-silvered mirror) which splits the photon state into a quantum superposition of the two considered earlier, so that its state now becomes a superposition of being transmitted, and aimed at the brown object, and of being reflected, and aimed at the green object. Quantum linearity demands that the two superposed photon states, as they emerge from the beam splitter each individually maintains the evolution that it would have achieved in the absence of the other. If quantum linearity holds universally, then the results of the photon impacts (for a photon of high energy) could be perfectly evidently macroscopic events.

This issue was clearly appreciated by Schrödinger, as he illustrated the problem with his famous "cat" thought experiment. In his version he considered a decay of

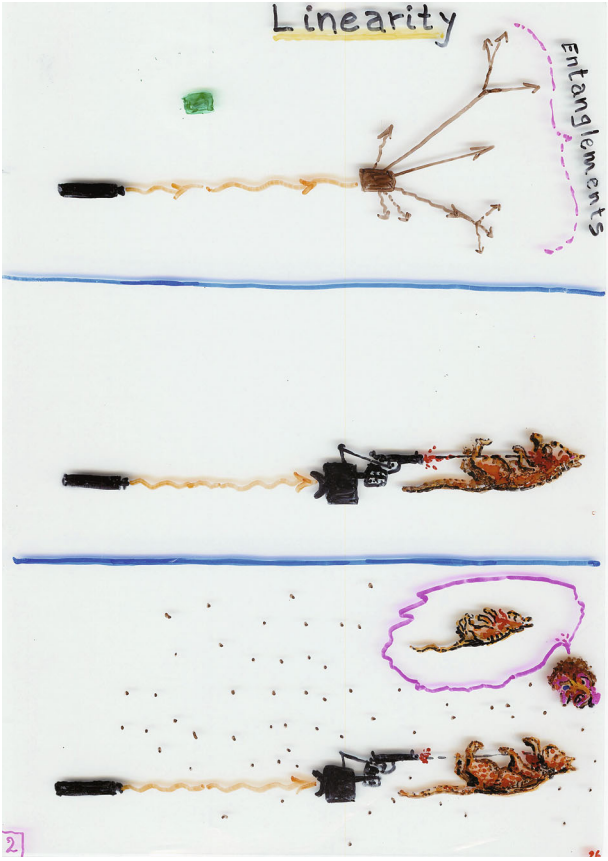



**Fig. 1** *Top* Single direct photon hits brown object emitting many particles. *Middle* Direct photon kills cat. *Bottom* Killed cat with environment, sad observer, sad thoughts

a radio-active atom, but it will fit in better with what I have to say if I employ a beam-splitter instead. In Fig. 1, the brown object is replaced by a murderous device (photon detector connected to a gun) which kills a cat. In Fig. 2, a mirror is inserted between the photon source and the device, so the cat survives. But in Fig. 3, the mirror is replaced by a beam-splitter so, according to quantum linearity (as demanded by **U**), the resulting quantum state is one involving a superposition of a dead and a live cat!

Sometimes people object to this direct kind of description of a quantum **U**-evolution, pointing out that I do not appear to have taken into consideration the complicated *environment* that must be entangled with the detector and the cat, etc.—since in the "environmental decoherence" description of quantum measurement, this environmental involvement is all-important, the unobserved degrees of freedom in the air molecules, for example, having to be averaged out in our quantum descriptions. Other objections come from the fact that no "observer" has been brought into the picture, it being frequently argued that it is by bringing in an observer's consciously perceived alternatives—necessarily of one clear outcome or another—that finally breaks the

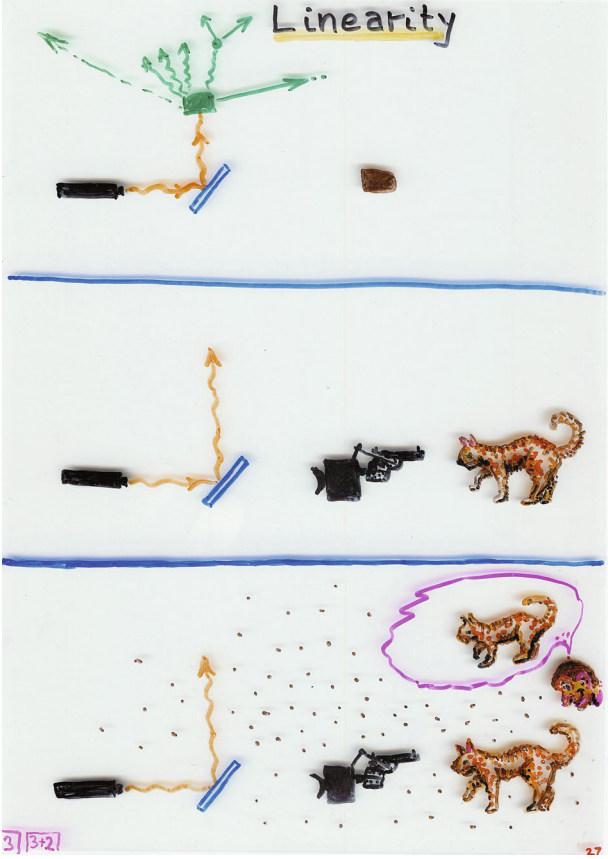



**Fig. 2** *Top* Photon reflected, hits green object emitting different particles. *Middle* Photon reflected, cat lives. *Bottom* Live cat, environment, happy observer, happy thoughts

chain of continuing superpositions. In Fig. 1, I have indicated the environment, with some dots to represent air molecules, in the situation described previously in Fig. 1. I have also included a human observer. The observer's conscious perception of the dead cat is depicted in a "thought bubble", but whether or not such a thing can be represented as part of the quantum state is not important here, because one can see from the observer's unhappy expression that the cat's demise has been consciously perceived. In Fig. 2, I have done the same, but with the mirror inserted, so that the cat remains alive, accompanied by a different environment, represented by some slightly differently placed dots. Now the observer consciously perceives the cat to be alive (as indicated in the thought bubble), and this is made evident from the now happy expression on the observer's face. None of this appears to affect the situation arising when the mirror is replaced by a beam splitter, as depicted in Fig. 3. The environment is now a superposition of these two possibilities, and the observer's facial expression is now a superposition of a smile and a frown, as the outward expression of the superposed conscious perception of a dead and live cat.

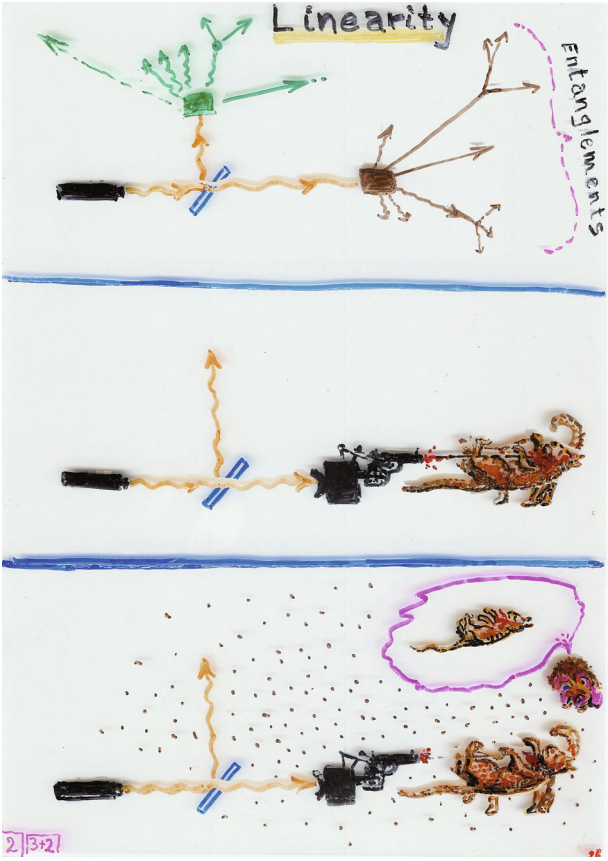



**Fig. 3** *Top* Beam-split photon produces both sets of particles in superposition. *Middle* Beam-split photon creates Schrödinger's cat. *Bottom* Schrödinger's cat with superposed environment, observer, and thoughts

Of course, some might object that conscious perceptions are not like this, and our normal streams of consciousness do not involve us in perceiving such gross superpositions. Of course they do not, in our actual experiences, but I see no contradiction in this kind of conscious perception, and we need to explain *why* they do not occur. The **U**-evolution of the standard quantum formalism, with its implied linearity, provides, in itself, no explanation for this remarkable but utterly familiar fact.

To get the right answer for the un-superposed nature of the perceived macroscopic world, we need to wheel in the *other* part of quantum formalism, namely the reduction **R** of the quantum state, according to which the state is taken as "jumping", probabilistically, to an eigenstate of some quantum operator that is taken to represent the action of quantum measurement. If we don't bring in **R** then we just don't get a description of the physical world that we all perceive. Strictly speaking, **R** is simply *inconsistent* with **U**. Whereas **U** is a continuous and deterministic evolution, **R** is (normally) discontinuous and probabilistic. Moreover, any measuring apparatus is itself constructed from the same kind of "stuff" that constitutes all the quantum systems that are under examination, being built from quantum particles, quantum fields and quantum forces.

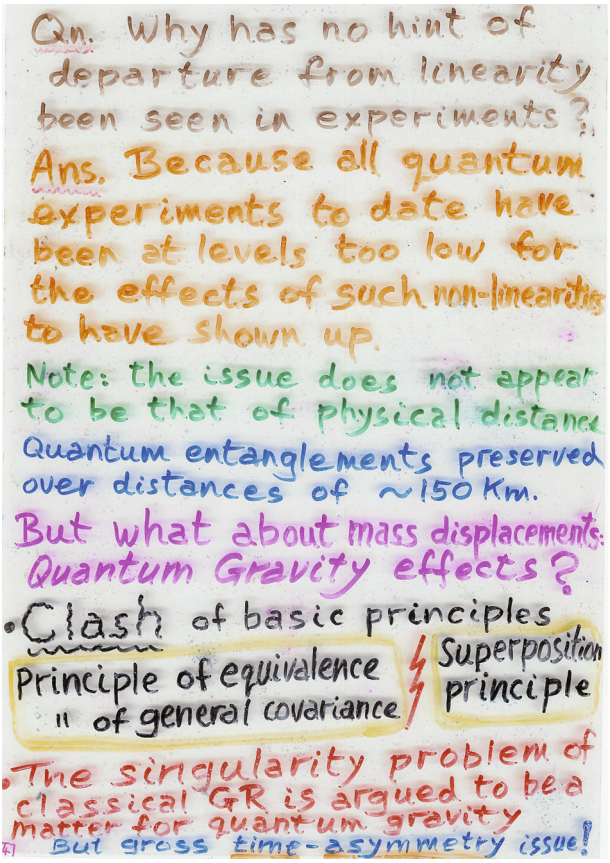



**Fig. 4** Clash between basic QM and GR principles

So why do we not use **U** to describe its action in the course of performing a measurement?

Of course there are innumerable different ways that quantum physicists and philosophers try patch up this inconsistency, normally taking **U** to represent an underlying truth of Nature and regarding **R** as coming about somehow as an approximation, or as an "interpretation" of how the quantum state is supposed to be viewed in relation to physical reality. I do not propose to enter into a discussion of any of these alternative viewpoints, but I merely present my own view, which is to regard **U** itself as an approximation to some yet-undiscovered non-linear theory. That theory would have to yield **R** as another approximation to a reality, describing how quantum measuring devices actually evolve in a way that subtly deviates from the action of **U**.

Again, there are numerous different approaches to such "realistic" evolutions, and I shall restrict attention to my own point of view [11–13], which is closely similar to one put forward, rather earlier (but motivated by somewhat different considerations), by Diósi [3,4], namely to regard *gravitation* as being the key to the issue of how standard quantum theory is to be modified. Other rather different gravitational proposals have

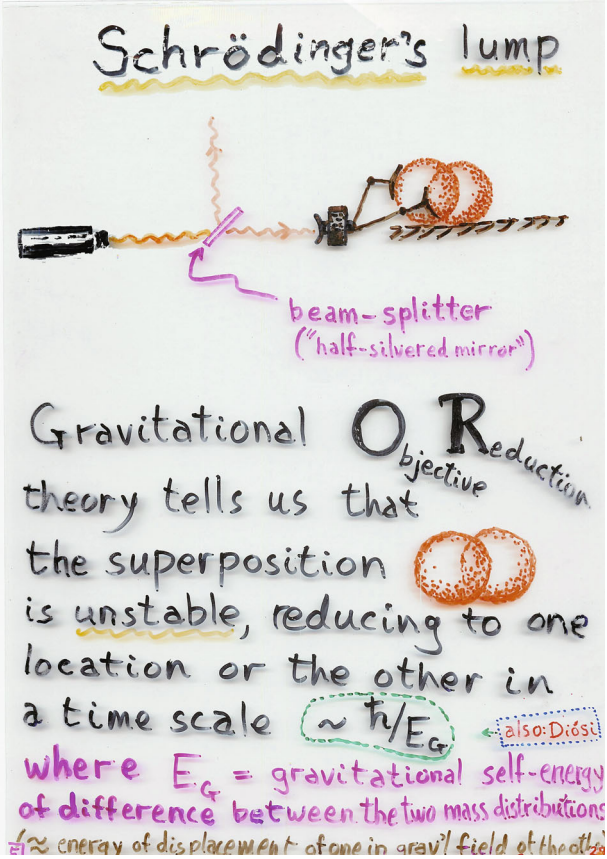



**Fig. 5** Schrödinger's lump; proposal for superposition decay

been proposed by [5–8, 17]. My own approach to this is to take the principles of general relativity as having a defining role in how quantum mechanics is to be modified. Some of the relevant issues, relating to the idea that gravity should be involved in this modification of quantum mechanics are listed in Fig. 4.

In Fig. 5, I have schematically depicted a more inanimate version of the Shrodinger cat (which had been illustrated previously in Fig. 3). Instead of the cat I have placed a simple massive (spherical) lump of material, which is moved to the right if the detector receives a photon but remains in its original location if that photon is deflected away from the detector, this photon having been fired from the laser on the left of the picture, and where there is to be a beam-splitter placed between the laser and the detector, all as before. The scheme that Diósi and I have proposed is concerned with a superposition between two states, each of which—had it been on its own—would have to be a *stationary* state. This superposition is to be taken as being *unstable* (owing to some non-linear effects coming from the putative unknown "gravitization" of standard quantum mechanics), with a lifetime $\tau$, of the order of

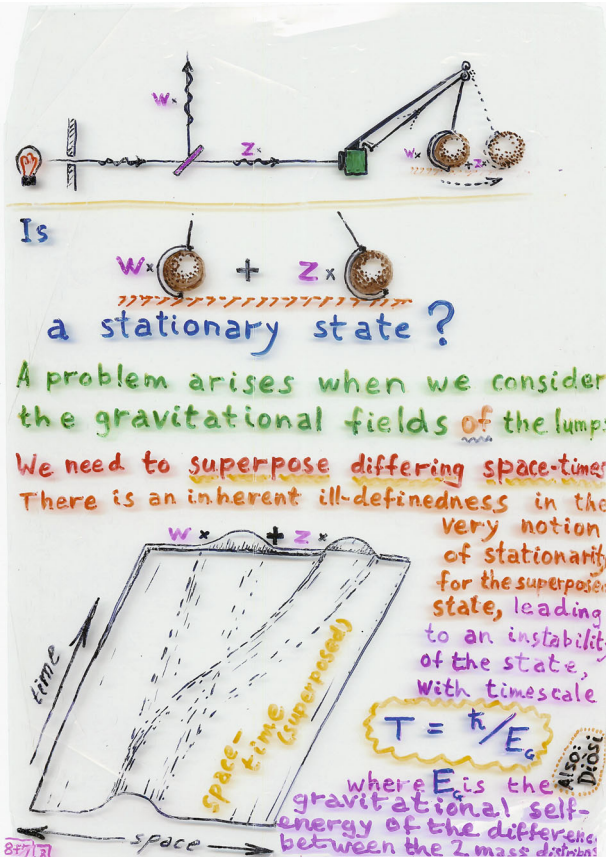



**Fig. 6** Schrödinger's lump gives space-time bifurcation

$$\tau \approx \hbar / E_G$$

where the quantity $E_G$ is taken as some fundamental *uncertainty* in the energy of the superposed state see [12], and the above formula is taken to be an expression of the Heisenberg time-energy uncertainty relation (in analogy with the formula relating the lifetime of a radioactive nucleus to its mass/energy uncertainty).

The quantity $E_G$ is the gravitational self-energy of the *difference* between the mass (expectation) distributions of the two stationary states in superposition. (If the two states merely differ from one another by a rigid translation, then we can calculate $E_G$ as the gravitational interaction energy, namely the energy it would cost to separate two copies of the lump, initially considered to be coincident and then moved to their separated locations in the superposition.) The calculation of $E_G$ is carried out entirely within the framework of Newtonian mechanics, as we are considering the masses involved as being rather small and moved very slowly, so that general-relativistic corrections can be ignored.

Nevertheless, we are to consider that regarding $E_G$ as an energy uncertainty comes from considerations of general-relativistic principles. In Fig. 6, I have schematically

**Fig. 7** Space-time bifurcation and one branch dies

illustrated the two space-times involved, being taken as more-or-less identical until the lump locations begin to separate. When the separation becomes significant, in an appropriate quantum-gravity sense, then one branch dies off and the other survives (Fig. 7). The time-scale for this to happen (according to the $\tau \approx \hbar/E_G$ formula) is estimated as being a measure of the amount of space-time difference (shown shaded in Fig. 8) in the regions where the two space-times in superposition differ from one another by an amount of order unity when measured in Planck units (i.e. taking $c = \hbar = G$=1). Thus the instability that results in just one of the space-time branches surviving sets in at the kind of scale when the space-time difference is of order unity. This measure of space-time difference is calculated according to the standard symplectic measure that can be estimated by considerations of linearized gravitational theory see [11].

Another, rather more rigorous, way of obtaining the $\tau \approx \hbar/E_G$ estimate is to appeal to the *principle of equivalence* see [15] which is the basic foundational principle of general relativity. Let us first imagine a table-top quantum experiment being performed, and we wish to calculate the wavefunction of the quantum system under

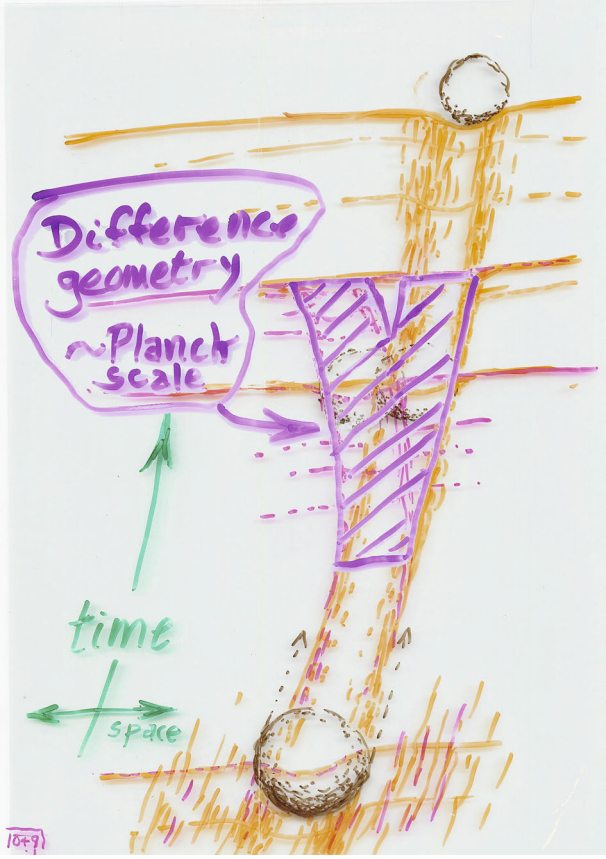



**Fig. 8** Decay after Planck-scale difference geometry measure

consideration, so that Schrödinger's equation can then be used to calculate the time-evolution of this state. In this particular case, we are interested in taking the Earth's *gravitational field* into consideration. We may envisage two different procedures for doing this (illustrated in Fig. 9). The first (with the green coordinates) would be the conventional quantum physicist's approach, where we simply include a term in the quantum Hamiltonian representing the Newtonian potential of the gravitational field, where we are just treating the gravitational field as providing an "ordinary force" in the same way that we would for any other force (such as an electric or magnetic force). This procedure gives us what I shall call the *Newtonian* wavefunction $\psi$ (written in green in Fig. 9). The second procedure is more in the spirit of Einstein's general relativity, where we imagine doing our quantum mechanics in a freely falling frame (purple coordinates), and in this frame there is no gravitational field, the accelerated fall cancelling it out, and our *Einsteinian* wavefunction $\Psi$ (written in purple in Fig. 9), described in terms of these free-fall coordinates, evolves according to a Schrödinger equation whose quantum Hamiltonian has no term representing the gravitational field.

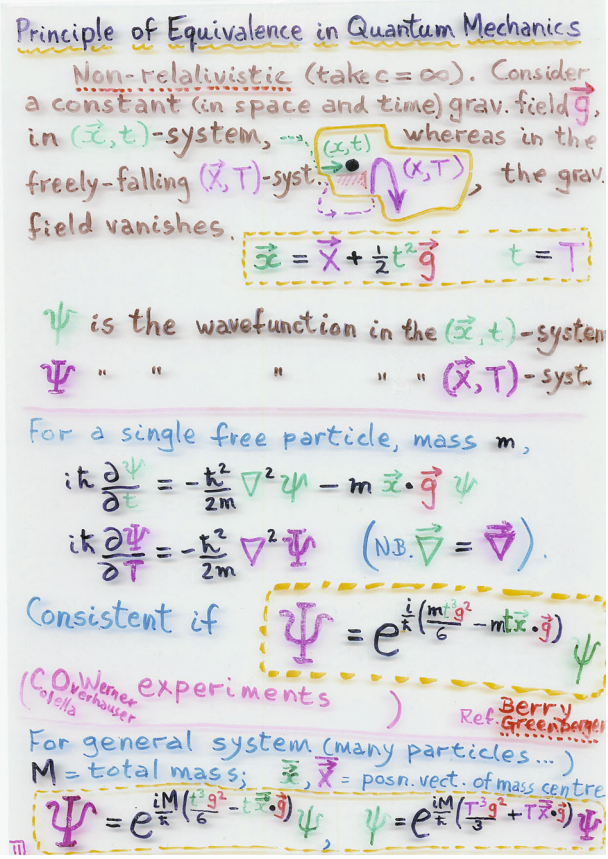



**Fig. 9** Principle of equivalence leads to phase difference

We would expect these two quantum evolutions to agree with one another—and, indeed a famous experiment performed in 1975, by Colella, Overhauser and Werner did confirm that Einstein's principle of equivalence is indeed respected by quantum mechanics in this respect. In fact, a direct calculation shows that the Newtonian and Einsteinian wavefunctions are related to each other simply by one being a *phase multiple* of the other. This multiple is explicitly shown in Fig. 9 (outlined with a yellow broken line), 3/4 of the way down the picture for the dynamics of a single particle, and at the bottom of the picture for the dynamics of a system of many particles.

Now, surely, this shows that standard quantum mechanics is perfectly consistent with Einstein's equivalence principle, since an overall *phase* multiplier for the wavefunction ought not to affect the interpretation in terms of probabilities, when a measurement (**R**) is performed on the system. However, there is an additional subtlety here, because if we examine this particular phase factor we see that it contains a term in $t^3$ in the exponent (with coefficient proportional the square $\mathbf{g}^2$ of the gravitational acceleration vector $\overrightarrow{g}$) where $t$(=T) is the time in both coordinate systems. Because of this non-linear dependence of the phase factor on $t$, we find that the notion of *positive frequency* differs in the two viewpoints (Newtonian and Einsteinian). Accordingly, the

two different reference systems, the Newtonian one and the Einsteinian one, refer to two different quantum field theory *vacua*.

The fact that physics in an accelerating frame is described by a different vacuum from that in an unaccelerated one is familiar in relativity, where according to what is called the "Unruh effect" [2,18] the accelerated vacuum is a *thermal vacuum* with a temperature $\hbar g/2\pi ck$. ($k$ being Boltzmann's constant, etc.). Here, we are considering the Galilean limit $c \to \infty$, so we see that the temperature goes to zero. Nevertheless, the accelerated vacuum remains different from the unaccelerated one even in this limit, although in this limiting case the accelerated vacuum is not a thermal one, but the phase discrepancy between the Newtonian and Einsteinian wavefunctions agrees with that shown in Fig. 9 (confirmed in a private communication from B. S. Kay).

Although, strictly speaking, the notion of alternative vacua is a feature of quantum *field* theory, rather than of the "Galilean" non-relativistic quantum mechanics that is under consideration here, the issue has direct relevance to the latter also. Standard quantum mechanics indeed requires that energies remain positive (i.e. that frequencies remain positive). This is not normally a problem, because the quantum dynamics governed by a positive-definite Hamiltonian will preserve this condition. But in the situation arising here we appear to be forced into the prospect of its violation unless the vacua are kept separate, the presence of the $t^3$ term coming about through the adoption of the Einsteinian perspective with regrd to the gravitational field.

In the particular experimental situation just considered, where the gravitational field is external and *given*, this difference in vacua is of no particular concern, since we can use one or the other description consistently without any discrepancy between the two choices of vacuum, as regards the results of measurement. But when the gravitational field itself becomes involved in a quantum superposition, the issue is more serious. In Fig. 10 I have illustrated this kind of situation, where a massive sphere is placed in a quantum linear superposition of two separate locations. We are to consider how quantum mechanics is to be used to treat this situation in accordance with the (preferred) *Einsteinian* viewpoint. The problem now is that in order to describe this superposition of gravitational fields, according to this viewpoint, we are confronted with the problem of superposing states that relate to two different vacua.

Strictly speaking this is *illegal*, according to the rules of quantum field theory, as the two states that are proposed to be under superposition belong to different Hilbert spaces. The general problem that arises would be that when we try to form an inner product between a state described in relation to one vacuum with a state described in relation to a different vacuum, we tend to get a *divergence* which violates the very rules that Hilbert spaces are supposed to satisfy (that inner products should be finite numbers), and for which the probability interpretation (Born rule) now makes nonsense. Nevertheless, we might take the view that such things come about merely because some mathematical idealization has been made (such as is the case for momentum states, for example, which are, strictly, not Hilbert space members) and that if full details of the situation are properly encoded, then such divergences would not occur. But how are we to deal with this when the very Hilbert spaces are diiferent, so that superpositions cannot actually be performed? An analogous situation might come from some mathematical treatments of, for example, a superconductor, where one might consider the superconducting state to be described in relation to a vacuum different

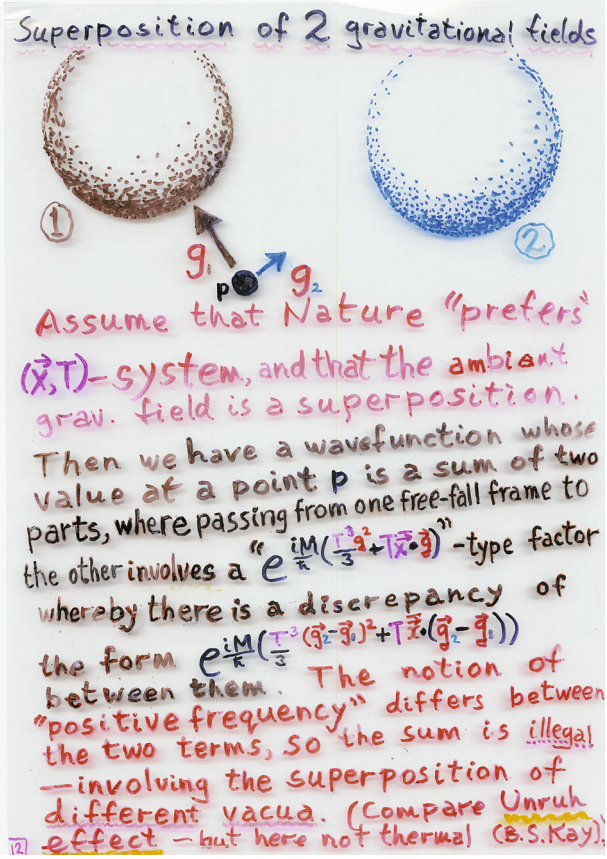



**Fig. 10** Displaced mass: illegal superposition of different vacua

from the standard one. Of course, if an actual physical superconductor were in a quantum state belonging to a Hilbert space with a different vacuum from that of the Hilbert space describing the system before the superconducting state were set up, then there would be no way to build a superconductor, since transitions from one Hilbert space to another, each with a different vacuum, would be "against the rules" of standard quntum field theory.

In the case of our superposed states of displaced massive spheres, we must imagine that the "illegality" in performing the superposition between states from different Hilbert states would take some time ($\approx \tau$) to make itself manifest in the physics. It has been hard, so far, to see how to make this completely unambiguous, but it seems that one can make a preliminary estimate by applying the Heisenberg time-energy uncertainty principle, rather in the way that this is done for a radioactive nucleus. For such a nucleus, there is a (mass-energy) uncertainty that is in reciprocal relation to its lifetime. Here the argument is similar (the converse of this), where we propose to estimate a superposition's lifetime from a fundamental energy uncertainty that it necessarily possesses by virtue of the clash between the general-relativistic and

quantum principles that are being invoked in relation to the description of stationary gravitational fields in quantum linear superposition. We can try to estimate the *error* that is involved in trying to form a stationary state out of the superposition of locations, where we take the view that the divergences involved in the superposition will come about from the coefficient of $t^3$in the "discrepancy" phase factor that we see in Fig. 10, namely

$$(M/3\hbar)\,(\vec{g}_2 - \vec{g}_1)^2.$$

If we take the integral of this expression over space, at one moment of the time $t(=T)$, as a measure of this error, then we obtain the quantity $E_G$ described earlier (the gravitational self-energy of the *difference* between the two mass distributions involved. (See [12] for this derivation and for an alternative motivation for this energy uncertainty.) When $E_G$ is substituted into Heisenberg's time-energy uncertainty formula, we get the aforementioned estimate of the lifetime of such a superposed state, before it spontaneously decays into one alternative or the other.

One point should be made clear about this proposal [3,4,11,12], namely that *all* **R** processes of state-reduction are to be the result of actions of this gravitational proposal. In a deliberate measurement by means of a conventional quantum detector, **R** would be triggered by some movement of mass within the detector, sufficient to reduce the state within a very small period of time $\sim \tau$. In other situations, where there is no deliberate measurement of the quantum system, but some considerable entanglement with an extended random environment, the major mass displacement would likely come about from the mass displacement of material within that environment, so when the environment state collapses to one or the other so does that of the quantum system itself, because the two parts, being entangled, both reduce together. This makes contact with the point of view whereby **R** is deemed to occur via *environmental decoherence*, but now we have a clear ontology of an objectively *real* reduction of the quantum state, and we do not have to appeal to some vague concept of the environmental degrees of freedom being, in some sense, "unobservable".

In order to get an impression of the role of $E_G$ we can first think of the case of a solid spherical ball of radius $a$, made from some uniform massive material of total mass $m$, where the ball is in a superposition of two locations, differing by a distance $b$. See Fig, 11. We use a Newtonian calculation to work out the value of $E_G$, obtaining the expression depicted in the figure. The only point of particular relevance here, in this expression, is the fact that for a displacement successively from coincidence to *contact* of the two instance of the sphere, the value of $E_G$ is already nearly 2/3 of the value it would reach for a displacement all the way out to infinity. Thus, for a uniformly solid body like this, we do not gain much by considering displacements in which the two instances of the body are moved apart by a distance of more than roughly its own diameter.

Of course, ordinary material is not really uniform in mass distribution, like this, but consists of huge numbers of nuclei where most of the mass is concentrated, where we might try to consider that the actual mass distribution is distributed among precise point-like locations of the quarks and electrons—marked in brown, in the middle of Fig. 12. However, when we then try to calculate $E_G$ by considering a very slightly

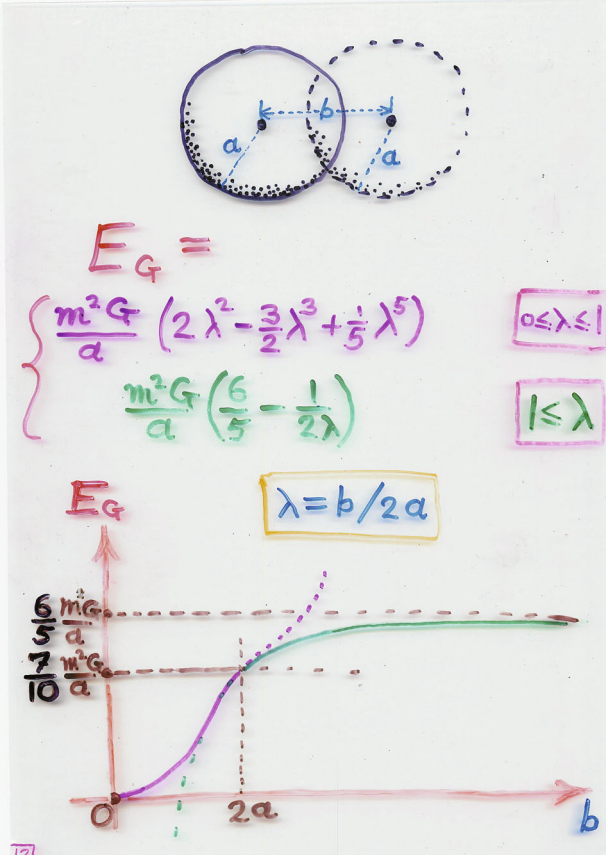



**Fig. 11** Newtonian self-energy for displaced uniform balls

shifted mass distribution, taking the *difference* between the two (the mass difference distiribution indicated by the brown–green difference shown in Fig. 13), then we get an infinite answer if the mass distributions are actually taken as the delta functions representing these point-particle locations. Such an infinite answer for $E_G$ would give a *zero* decay time, in other words an instantaneous reduction of any superposed state of actual material—and we would have no quantum mechanics!

Clearly this is not the correct answer, and indeed the procedure has not been carried out correctly, because (as noted earlier) the procedure of calculating $E_G$ applies only when each of the two states in superposition is actually *stationary*. A state involving delta function mass distributions like this would not be stationary, according to Schrödinger's equation, since the delta-function mass distributions would instantly spread out from these locations. What we need to do is first to solve the stationary Schrödinger equation for the material in our object. Even this would lead to a nonsense, however, if carried out strictly, for an actual stationary state, since such a state would have to be spread out over the entire universe, and we would now get the opposite problem of obtaining a *zero* answer for $E_G$.

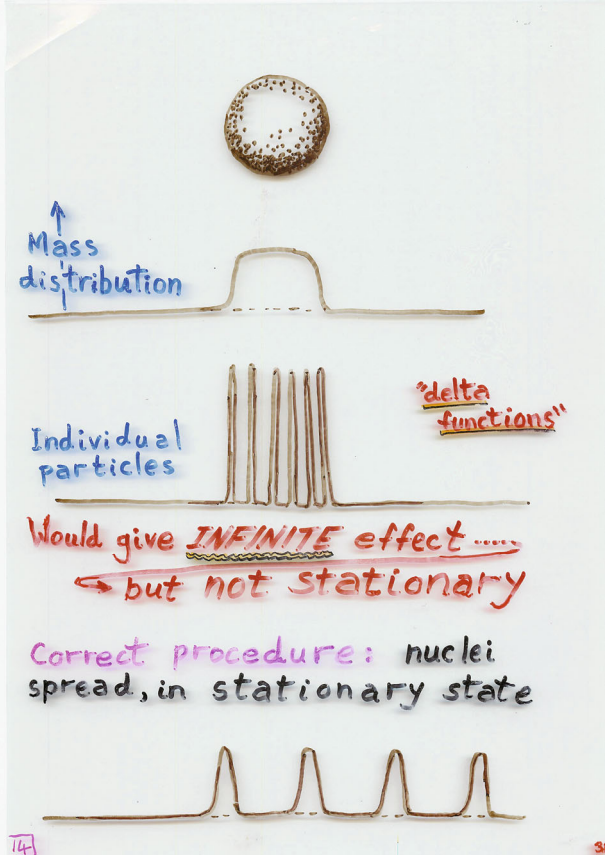



**Fig. 12** Different models of mass concentration

What one tends to do, in practice, when considering the wavefunction of a stationary state is to factor out the mass centre, regarding the wavefunction to be concerned with distances *relative* to the mass centre. This procedure is a little artificial, and one can adopt the more systematic procedure of modifying the Schrödinger equation in the form of the self-coupled "Schrödinger-Newton equation", for a wavefunction $\psi$, which incorporates an additional term in the Hamiltonian, that is given by the gravitational potential due to the expectation value of the mass distribution in $\psi$ itself see [10,13]. The "Newton" term prevents the stationary solutions from being infinitely spread out. In situations of relevance here, to a reasonable enough approximation, this amounts to the aforementioned procedure of factoring out the mass centre, and then just solving the Schrödinger equation. This saves us from having to solve the non-linear Schrödinger-Newton equation in detail in order to calculate $E_G$. On the other hand, incorporating the Schrödinger-Newton equation has another theoretical importance in also providing the possible stable stationary states to which a quantum superposition can reduce by the operation of **R**.

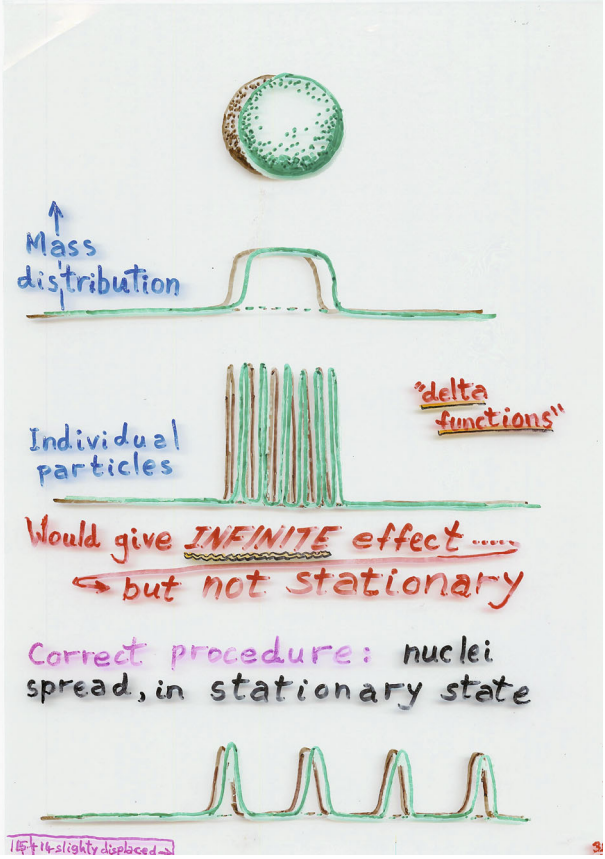



**Fig. 13** Different models of displaced mass concentrations

It might be presumed, since we are looking at space-time differences whose measure is of the order of the extremely tiny Planck scale, that nothing of observable consequence would be likely to come from considertions of this kind. Indeed, experiments aimed at testing the influences of quantum mechanics on space-time structure—the presumed consequences of "quantum gravity" in the usual interpretation of that term—would be enormously far from what can be achieved by current technology. However, we are here concerned with the role of gravitational principles on the structure of quantum mechanics, not the other way about, where the experimental situation is very different. One way of looking at this issue is to realize that the scales at which quantum mechanics is expected to influence space-time structure would indeed be the absurdly tiny Planck length $(\hbar G/c^3)^{1/2} \approx 1.6 \times 10^{-35}$m and Planck time $(\hbar G/c^3)^{1/2} \approx 5.4 \times 10^{-44}$s, these being so very tiny, compared with ordinary scales, because they involve dividing a very small quantity by a very large one. However, the quantity of relevance to our considerations of the "gravitization of quantum mechanics" under consideration here is $\tau \approx \hbar/E_G$, which is one very small quantity divided by another very small quantity (since $E_G$is scaled by the gravitational constant), which

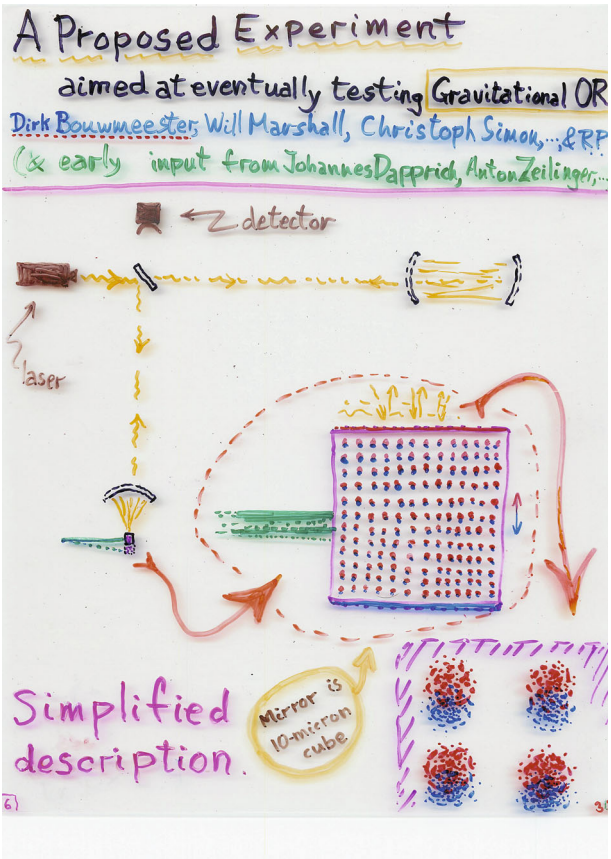



**Fig. 14** Bouwmeester experiment (schematic)

need be neither very small nor very large. Accordingly, one must examine the quantities involved in any proposed experiment very carefully, to see whether the time scale $\tau$ turns out to be one that comes within the scope of the experiment being envisaged.

In Fig. 14, I have sketched the rough idea see [9] underlying an experiment that has been under development by Dirk Bouwmeester and colleagues, at the Universities of Leiden (Netherlands) and California (Santa Barbara, CA, USA), for several years now, where a tiny cubical mirror ($\sim 10^{-5}$m in dimension) is to be put into a superposition of two different locations, differing from one another by about the diameter of an atomic nucleus. This is to be achieved by multiple impacts ($\sim 10^6$ of them) by a single photon, which has been beam-split into a quantum superposition of two alternative routes, in only one of which the photon hits the tiny mirror. This mirror is to be maintained in its superposed state for seconds or perhaps up to minutes which, depending on the details, might be a sufficient time $\tau$ for the decay of the superposition to take place (according to the aforementioned state-reduction proposal). This spontaneous state reduction could be tested by subsequently reversing the photon routes, and observing whether there is consequent lack of phase coherence between the two branches of the

beam-split photon, such breaking of phase coherence leading to the possible activation of the detector at the top of the picture. Clearly there are enormous difficulties in eliminating unwanted possibilities of decoherence, etc., but there appears to be a reasonable prospect of a result from experiments of this kind within the next decade.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

## References

1. Colella, R., Overhause, A.W., Werner, S.A.: Observation of gravitationally induced quantum interference. Phys. Rev. Lett. **34**, 1472 (1975)
2. Davies, P.C.W.: Scalar production in Schwarzschild and Rindler metrics. J. Phys. A. **8**, 609 (1975)
3. Diósi, L.: A universal master equation for the gravitational violation of quantum mechanics. Phys. Lett. **120A**, 377–81 (1987)
4. Diósi, L.: Models for universal reduction of macroscopic quantum fluctuations. Phys. Rev. **A40**, 1165–1174 (1989)
5. Károlyházy, F.: Gravitation and quantum mechanics of macroscopic bodies. Nuovo Cim. A **42**, 390 (1966)
6. Károlyházy, F.: Gravitation and quantum mechanics of macroscopic bodies. Magyar Fizikai Polyoirat **12**, 24 (1974)
7. Károlyházy, F., Frenkel, A., Lukács, B.: On the possible role of gravity on the reduction of the wave function. In: Penrose, R., Isham, C.J. (eds.) Quantum Concepts in Space and Time, pp. 109–128. Oxford University Press, Oxford (1986)
8. Kibble, T.W.B.: Is a semi-classical theory of gravity viable? In: Isham, C.J., Penrose, R., Sciama, D.W. (eds.) Quantum Gravity 2: A Second Oxford Symposium, pp. 63–80. Oxford University Press, Oxford (1981)
9. Marshall, W., Simon, C., Penrose, R., Bouwmeester, D.: Towards quantum superpositions of a mirror. Phys. Rev. Lett. **91**(13–16), 130401 (2003)
10. Moroz, I.M., Penrose, R., and Tod, K.P.: Spherically-symmetric solutions of the Schrödinger-Newton equations. Class. Quantum Grav. **15**, 2733–2742 (1998)
11. Penrose, R.: Gravity and quantum mechanics. In: Gleiser, R.J., Kozameh, C.N., Moreschi, O.M. (eds.) General Relativity and Gravitation 13. Part 1: Plenary Lectures 1992, pp. 179–189. Institute of Physics Publishing, Bristol, Philadelphia (1998).
12. Penrose, R.: On gravity's role in quantum state reduction. Gen. Relativ. Gravit. **28**, 581–600 (1996)
13. Penrose, R.: Quantum computation, entanglement and state-reduction. Phil. Trans. Roy Soc. (Lond.) A **356**, 1927–1939 (1998)
14. Penrose, R.: The Road to Reality: A Complete Guide to the Laws of the Universe. Jonathan Cape, London (2004)
15. Penrose, R.: Black holes, quantum theory and cosmology (Fourth International Workshop DICE 2008). J. Phys. Conf. Ser. **174**, 012001 (2009)
16. Penrose, R.: On the gravitization of quantum mechanics 2: conformal cyclic cosmology. Found. Phys. (2014). doi:10.1007/s10701-013-9763-z
17. Percival, I.C.: Quantum spacetime fluctuations and primary state diffusion. Proc. R. Soc. Lond. **A451**, 503–13 (1995)
18. Unruh, W.G.: Notes on black hole evaporation. Phys. Rev. D **14**, 870 (1976)

# BEFORE THE BIG BANG: AN OUTRAGEOUS NEW PERSPECTIVE AND ITS IMPLICATIONS FOR PARTICLE PHYSICS

Roger Penrose
Mathematical Institute, 24-29 St Giles', Oxford OX1 3LB, U.K.

## THE BASIC CONUNDRUM

Proposals for describing the initial state of the universe hardly ever address a certain fundamental conundrum [1] — yet this is a conundrum whose significance is, in a certain sense, obvious. The issue arises from one of the most fundamental principles of physics: the *Second Law of thermodynamics*. According to the Second Law, roughly speaking, the entropy of the universe increases with time, where the term "entropy" refers to an appropriate measure of *disorder* or lack of "specialness" of the state of the universe. Since the entropy increases in the future direction of time, it must decrease in the past time-direction. Accordingly, the initial state of the universe must be the most special of all, so any proposal for the actual nature of this initial state must account for its extreme specialness. Proposals have been put forward from time to time (such as in various forms of "inflationary cosmology" and the previously popular "chaotic cosmology") in which it is suggested that the initial state of the universe ought to have been in some sense "random", and various physical processes are invoked in order to provide mechanisms whereby the universe might be driven into the special state in which it appears actually to have been in, at slightly later stages. But "random" means "non-special" in the extreme; hence the conundrum just referred to.

Sometimes theorists have tried to find an explanation via the fact that the early universe was very "small", this smallness perhaps allowing only a tiny number of alternative initial states, or perhaps they try to take refuge in the *anthropic principle*, which would be a selection principle in favour of certain special initial states that allow the eventual evolution of intelligent life. Neither of these suggested explanations gets close to resolving the issue, however. It may be seen that, with time-symmetrical dynamical laws, the mere smallness of the early universe does not provide a restriction on its degrees of freedom. For we may contemplate a universe model in the final stages of collapse. It must do *something*, in accordance with its dynamical laws, and we expect it to collapse to some sort of complicated space-time singularity, a singularity encompassing as many degrees of freedom as were already present in its earlier non-singular collapsing phase. Time-reversing this situation, we see that an initial singular state could also contain as many degrees of freedom as such a collapsing one. But in our actual universe, almost all of those degrees of freedom were somehow not activated.

What about the anthropic principle? Again, this is virtually no help to us whatever in resolving our conundrum. It is normally assumed that life had to arise via complicated evolutionary processes, and these processes required particular conditions, and particular physical laws, including the Second Law. The Second Law was certainly a crucial part of evolution, in the way that our particular form of life actually came about. But the very action of this Second Law tells us that however special the universe may be now, with life existing in it now, it must have been far *more* special at an earlier stage in which life was not present. From the purely anthropic point of view, this earlier far more special phase was not needed; it would have been much more likely that our present "improbable" stage came about simply by chance, rather than coming about via an earlier even more improbable stage. When the Second Law is a crucial component, there is always a far more probable set of initial conditions that would lead to this same state of affairs, namely one in which the Second Law was *violated* prior to the situation now!

As another aspect of this same issue, we may think of the vastness of our actual universe, most of which had no actual bearing on our existence. Though very special initial conditions were indeed required for our existence in our particular spatial location, we did not actually need these same special conditions at distant places in the universe. Yet as we look out at the universe, we see the same kind of conditions, acting according to the same Second Law of thermodynamics, no matter how far out we look. If we take the view that the Second Law was introduced in our vicinity merely for our own benefit, then we are left with no explanation for the extravagance of this same Second Law having to be invoked uniformly throughout the universe, as it appears to be as far as our powerful instruments are able to probe.

## THE ENORMITY OF THE SPECIALNESS

In order to stress the extraordinary scale of this problem, and the intrinsic implausibility of explanations of this kind, it is helpful to enter a little more precisely into the definition of entropy, and to estimate the entropy magnitudes that we have to contend with. Boltzmann provided us with a beautiful formula for the entropy $S$ of a system:

$$S = k \log V.$$

Here $k$ is Boltzmann's constant and $V$ is the volume of a certain region in the total phase space $\boldsymbol{P}$ of the system under consideration. We are assuming $\boldsymbol{P}$ to be "coarse-grained" into sub-regions, each sub-region representing states that are deemed to be indistinguishable with regard

to any reasonable macroscopic parameter. (There is clearly an element of arbitrariness or subjectivity, here, as to which parameters are to be regarded as macroscopically discernible and which are deemed to be effectively "unmeasurable". In practice, there is a considerable robustness with regard to this arbitrariness, and it is reasonable to disregard this issue in the present discussion.) Any particular state of the system under consideration will be specified by some point $x$ of $\boldsymbol{P}$, and the quantity $V$ is then the volume of the particular sub-region of $\boldsymbol{P}$ which contains $x$.

With regard to future time-evolution of the system, the Second Law can be understood as the fact that, as the system evolves, the point $x$ moves within $\boldsymbol{P}$ so that with overwhelming probability it enters sub-regions of successively larger and larger volume $V$. This arises from the fact that, in practice, the sub-regions differ stupendously in size. The logarithm in Boltzmann's formula helps here (as does the smallness of $k$, in ordinary units), because there need only be a modest increase in $S$ when $x$ moves from one sub-region into a neighbouring one of stupendously larger volume. But this is only the easy half of our understanding of the Second Law. The difficult half is to understand why, when we *reverse time*, $x$ enters successively *tinier* sub-regions of $\boldsymbol{P}$. It does this because it has ultimately to reach the exceptionally tiny region $\boldsymbol{B}$ which represents the Big Bang itself. The difficult half of the Second Law involves an understanding of why the universe had to start off in such an extraordinarily special state. And to understand how special the Big Bang actually was, we need to compare the volume of $\boldsymbol{B}$ with that of the entire phase space $\boldsymbol{P}$.

One point of concern is the fact that the entire volume might be infinite, as it certainly would be in the case of a spatially infinite universe. This issue, while of relevance, is not of major importance for our considerations here. There is also the issue of how we get a finite phase-space volume when some of the parameters would be describing continuous fields. I shall evade this latter issue by assuming that it is dealt with by quantum mechanics, where for a finite universe of bounded energy content we may assume only finitely many quantum states.

To deal with a spatially infinite universe, I shall assume that we need consider only, say, that comoving portion of the universe that intersects our past light cone. This contains something of the order of $10^{80}$ baryons. To obtain a lower bound for the volume of $\boldsymbol{P}$, for this situation, we can consider the entropy that arises when this number of baryons is collapsed into a black hole. For this, we use the Bekenstein–Hawking entropy formula $S_{BH}=8\pi^2 kGm^2/hc$ for a spherical black hole of mass $m$ and find a value of the order of $10^{123}$. If we collapsed the dark matter also into this black hole, we would get a considerably larger entropy (and, for a continually expanding universe, we should consider even larger values than this), but this value represents a usable lower bound. Recalling the logarithm in Boltzmann's formula (a natural logarithm, but that is of no concern), we get that the volume of $\boldsymbol{P}$ is greater than that of $\boldsymbol{B}$ by a factor that exceeds

$$10^{10^{123}}.$$

This gives us some idea of the enormity of the precision in the Big Bang!

## THE GEOMETRIC NATURE OF THE SPECIALNESS

A seeming paradox arises from the fact that our best evidence for the very existence of the Big Bang arises from observations of the microwave background radiation—frequently referred to as the "flash of the Big Bang", greatly cooled down to its present value of ~2.7K. The intensity of this radiation, as a function of frequency, matches the Planck radiation formula extraordinarily closely, giving us impressive evidence of an early universe state with matter in *thermal equilibrium*. But thermal equilibrium is represented, in phase space $\boldsymbol{P}$, as the coarse-graining sub-region of largest volume (so large that it normally exceeds all others put together). This corresponds to *maximum* entropy, so we reasonably ask: how can this be consistent with the Second Law, according to which the universe started with a very *tiny* entropy?

The answer lies in the fact that the high entropy of the microwave background refers only to the *matter* content of the universe and not to the *gravitation field*, as would be encoded in its space-time geometry in accordance with Einstein's general relativity. What we find, in the early universe, is an extraordinary *uniformity*, and this can be interpreted as the gravitational degrees of freedom that are potentially available to the universe being not excited at all. As time progresses, the entropy rises as the initially uniform distribution of matter begins to clump, as the gravitational degrees of freedom begin to be taken up. This allows stars to be formed, which become much hotter than their surroundings (a thermal imbalance that all life on Earth depends upon), and finally this gravitational clumping leads to the presence of black holes (particularly the huge ones in galactic centres), which represent an enormous increase in entropy.

Although, in general, there is no clear geometric measure of the entropy in a gravitational field in general relativity, we can at least provide proposals for the *non-activation* of gravitational degrees of freedom at the Big Bang. I have referred to such a proposal as the *Weyl Curvature Hypothesis* (WCH) [2]. In Einstein's theory the *Ricci curvature* $R_{ab}$ is directly determined by the gravitational sources, via the energy-momentum tensor of matter (analogue of the charge-current vector $J^a$ in Maxwell's electromagnetic theory) and the remaining part of the space-time Riemann curvature, namely the *Weyl curvature* $C_{abcd}$, describes gravitational degrees of freedom (analogue of the field tensor $F_{ab}$ of Maxwell's theory). WCH—which is a time-asymmetrical hypothesis—asserts that *initial* space-time singularities

must be constrained to have $C_{abcd}$=0 (in some appropriate sense), whereas final space-time singularities (as occur inside black holes) are unconstrained.

What appears to be the most satisfactory form of WCH has been studied extensively by Paul Tod [3]. This proposes that an initial space-time singularity can always be represented as a smooth *past boundary* to the *conformal geometry* of space-time. In conformal geometry, we consider the space-time structure that is invariant under rescalings of the metric

$$g_{ab} \rightarrow \hat{g}_{ab} = \Omega^2 g_{ab}$$

where Ω is taken to be a smooth positive scalar field on space-time. Another way of specifying the conformal geometry of space-time is simply to take the family of *null cones* (in the tangent spaces of the space-time points) as defining the geometry. It may be noted that the conformal structure contains 9 out of the 10 components of the metric, the overall scale providing, in effect, the $10^{th}$.

Tod's formulation of WCH is the hypothesis that we can adjoin a (past-spacelike) hypersurface boundary to space-time in which the conformal geometry can be mathematically extended smoothly through it, to the past side of this boundary. This amounts to "stretching" the metric by a conformal factor Ω which becomes *infinite* at the Big Bang singularity, so that we get a smooth metric $\hat{g}_{ab}$ which actually extends across this boundary.

## CONFORMAL CYCLIC COSMOLOGY

So far, we regard the conformal "space-time" prior to the Big Bang as a mathematical fiction, introduced solely in order to formulate WCH in a mathematically neat way. However, my "outrageous" proposal [4] is to take this mathematical fiction seriously as something *physically real*. But what "physical reality" can we consistently attach to this space-time occurring "before the Big Bang"? As a clue to this possibility, we should consider the nature of the physics that is presumed to be taking place just *after* the Big Bang. (I am going to ignore the possibility of inflation here, and assume that such exponential expansion did not actually take place after the Big Bang. The issue of inflation, in relation to this scheme, is considered in the section on physical implications, below.) As we approach the Big Bang, moving back in time, we expect to find temperatures that are increasingly great. And the greater the temperature, the more irrelevant the rest masses of the particles involved will become, so these particles are effectively massless near the Big Bang. Now, massless particles (of whatever spin) satisfy conformally invariant equations [5]. I am going to suppose that the *interactions* between these massless entities are also described by conformally invariant equations. (This seems to be consistent with current understanding of particle physics.) With such conformal invariance holding in the very early universe, the universe has no way of "building a clock". So it loses track of the scaling which determines the full space-time metric, while retaining its conformal geometry.

We may apply considerations of this kind also to the distant future of the universe. If we assume that in the very remote future, conformally invariant equations again govern the universe's contents, then we can apply the same mathematical trick as before, but now in the reverse sense that we look for a boundary at which the conformal factor Ω becomes *zero*, rather than infinite. This amounts to using a metric, such as $\hat{g}_{ab}$ above, in which the future infinity is "squashed down" to be a finite boundary to space-time, which is conformally regular in the sense that the space-time can be mathematically extended across this future boundary as a smooth conformal manifold [5]. If we also assume that there is a positive cosmological constant present, as current observations appear to point strongly towards, then we find that this future conformal boundary is *spacelike*.

There is, however, a crucial difference between the use of a conformal boundary to study the future asymptotics of a space-time and Tod's use of a conformal boundary to treat the Big Bang. For in the latter case the very validity of this trick provides a formulation of WCH, whereas it the future situation of an expanding universe with conformally invariant contents, the validity of this procedure is more-or-less *automatic* [5]. Physically, we may think that again in the very remote future, the universe "forgets" time in the sense that there is no way to build a clock with just conformally invariant material. This is related to the fact that massless particles, in relativity theory, do not experience any passage of time. We might even say that to a massless particle, "eternity is no big deal". So the future boundary, to such an entity is just like anywhere else. With conformal invariance both in the remote future and at the Big-Bang origin, we can try to argue that the two situations are *physically identical*, so the remote future of one phase of the universe becomes the Big Bang of the next. This suggestion is my "outrageous" *conformal cyclic cosmology*" (CCC) [4].

## PHYSICAL IMPLICATIONS

There are certain important assumptions involved in CCC, in order that only conformally invariant entities survive to eternity . One of these is that black holes will all eventually evaporate away and disappear. This evaporation is a consequence of Stephen Hawking's quantum considerations, and these are now normally accepted. There is, however, the issue (connected with the so-called "information paradox") of whether they would actually ultimately disappear or leave some form of "remnant". I am here taking the more conventional view that they would indeed disappear in a final (cosmologically very mild) explosion. More serious, for CCC, is how to get rid of massive fermions and massive charged particles. It is not too unconventional to assume that protons will ultimately decay, or even that there is one variety of neutrino that is massless, but the real

problem lies with electrons. A good many of them will annihilate with positively charged particles, but there will be a relatively small number of "stray" charged particles which become trapped in their ultimate event horizons, being unable to come in contact with other particles of opposite charge. There are various possible ways out of this, none of which is part of conventional particle physics. One possibility is that electric charge is not exactly conserved, so that within the span of eternity, electric charge would eventually disappear. A much more satisfying possibility, from my own perspective, is that the electron's mass will eventually decay away—and, again, there is all of eternity for this to happen, so the possibility may not be too outrageous to contemplate.

This last possibility is tied up with the issue of the strength of the gravitational interaction, which I have postponed in my discussion here. In the background of conformal geometry, the strength of gravity may be considered as being infinitely large at the Big Bang (which is, in a sense, why the gravitational degrees of freedom must initially be set to zero), and this strength gets smaller as time progresses, eventually reducing to zero at the final boundary. To express all this in a satisfactory mathematical framework for CCC, we need to reformulate general relativity in a conformally invariant way. This can indeed be done. We take advantage of the fact that the Weyl tensor $C_{abc}{}^{d}$ is conformally invariant, and provides a precise measure of the conformal curvature of space-time. We can define the gravitational "spin-2 field" $K_{abc}{}^{d}$ to be described by $C_{abc}{}^{d}$ with respect to the original space-time metric $g_{ab}$, but when we pass to the conformally related metric $\hat{g}_{ab}=\Omega^2 g_{ab}$ we find that, curiously, $K_{abc}{}^{d}$ picks up a factor of $\Omega^{-1}$, which $C_{abc}{}^{d}$ does not [5].

This has the implication that gravitational radiation (described by $K_{abc}{}^{d}$) actually survives at the future boundary (whereas $C_{abc}{}^{d}$ vanishes there) and its presence shows up as a non-zero normal derivative of $C_{abc}{}^{d}$ at the boundary. This gives rise to density fluctuations at the Big bang, and possibly primordial gravitational radiation. The details of all this have yet to be worked out, but, in principle at least, there should be clear-cut predictions which should be observable. One important issue is how this compares with the detailed observations of temperature variations in the microwave background and the near scale-invariance of the initial density fluctuations. This scale invariance is normally taken as a success of inflationary theory. It will be interesting to see whether CCC leads to a similar implication with regard to these fluctuations, as it also involves an exponential expansion, though this occurs *before* the Big Bang in CCC, rather than afterwards.

## REFERENCES


[1] See, for example, Penrose, R. (2004) *The Road to Reality: A complete guide to the Laws of the Universe* (Jonathan Cape, London); Chapter 27.

[2] Penrose, R. (1979) Singularities and Time-Asymmetry, in *General Relativiy: an Einstein Centenary*, eds. S.W. Hawking and W. Israel (Cambridge University Press).

[3] Tod K.P. (2003) *Class. Quantum Grav.* **20** 521-534.

[4] Penrose, R (2005) http://www.newton.cam.ac.uk/webseminars/pg+ws/2005/gmr/ gmrw04/1107/ penrose/. Related ideas, but dependent upon extra space dimensions, were put forward earlier: Veneziano, G. (1998), see A Simple/Short Introduction to Pre-Big-Bang Physics/Cosmology, hep-th/9802057 v2 in the LANL Archive; also Steinhardt, P.J. and Turok N. (2001) A Cyclic Model of the Universe http://xxx.lanl.gov/PS_cache/hep-th/pdf/0111/0111030.pdf

[5] Penrose, R. (1965) Zero rest-mass fields including gravitation: asymptotic behaviour, *Proc. Roy. Soc. London*, **A284**, 159-203.

# *Space-time and Cosmology*

*ROGER PENROSE*

THE TANNER LECTURES ON HUMAN VALUES

Delivered at

Cambridge University
February 13–15, 1995

SIR ROGER PENROSE is Rouse Ball Professor of Mathematics at Oxford University, as well as Francis and Helen Pentz Distinguished Professor of Physics and Mathematics at Pennsylvania State University. He was educated at University College, London, and received his Ph.D. from Cambridge University. He is a fellow of the Royal Society and recipient of numerous prizes and awards, including the 1988 Wolf Prize that he shared with Steven Hawking for their understanding of the universe. His numerous publications include *Shadows of the Mind* (1994), *The Emperor's New Mind* (1994), *Spinors and Space Time* (1984 and 1986, coauthor), and *Techniques of Differential Topology* (1972). He was knighted in 1994 for services to science.

The title of these lectures is "The Large, the Small, and the Human Mind" and the subject of this first lecture is the Large.[1] The first and second lectures are concerned with our physical Universe, which I represent very schematically as the "sphere" in my first picture.

However, these will not be "botanical" lectures, telling you in detail what is here and what is there in our Universe, but rather I want to concentrate upon understanding of the actual laws that govern the way the world behaves. One of the reasons that I have chosen to divide my descriptions of the physical laws between two lectures, namely, the Large and the Small, is that the laws that govern the large-scale behaviour of the world and those that govern its small-scale behaviour seem to be very different. The fact that they seem to be so different, and what we might have to do about his seeming discrepancy, is central to the subject of the third lecture —which is where the human mind comes in.

Since I shall be talking about the physical world in terms of the physical theories that underlie its behaviour, I shall also have to say something about another world, the Platonic world of absolutes, in its particular role as the world of mathematical truth.

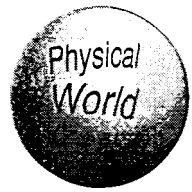



FIGURE 1.

[1] This is the first of three lectures with the overall title "The Large, the Small, and the Human Mind." All three lectures will be published by Cambridge University Press.

One can well take the view that the "Platonic world" contains other absolutes, such as the Good and the Beautiful, but I shall be concerned here only with the Platonic concepts of mathematics. Some people find it hard to conceive of this world as existing on its own. They may prefer to think of mathematical concepts merely as idealisations of our physical world —and, on this view, the mathematical world would be thought of as emerging from the world of physical objects (figure 2).

Now, this is not how I think of mathematics, nor, I believe, is it how most mathematicians or mathematical physicists think about the world. They think about it in a rather different way, as a structure precisely governed according to timeless mathematical laws. Thus, they prefer to think of the physical world, more appropriately, as emerging out of the ("timeless") world of mathematics, as illustrated in figure 3. This picture will have importance for what I shall say in the third lecture, and it also underlies most of what I shall say in the first two.

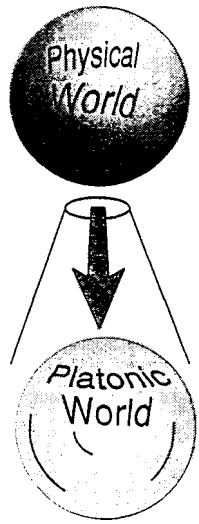



FIGURE 2.

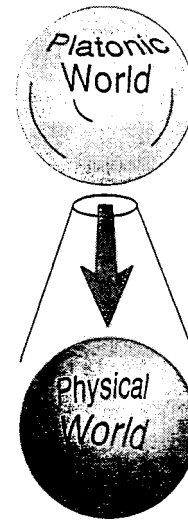



FIGURE 3.

One of the remarkable things about the behaviour of the world is how it seems to be grounded in mathematics to a quite extraordinary degree of accuracy. The more we understand about the physical world, and the deeper we probe into the laws of nature, the more it seems as though the physical world almost evaporates and we are left only with mathematics. The more deeply we understand the laws of physics, the more we are driven into this world of mathematics and of mathematical concepts.

Let us look at the scales we have to deal with in the Universe and also the role of our place in the Universe. I can summarise all these scales in a single diagram (figure 4). On the left-hand side

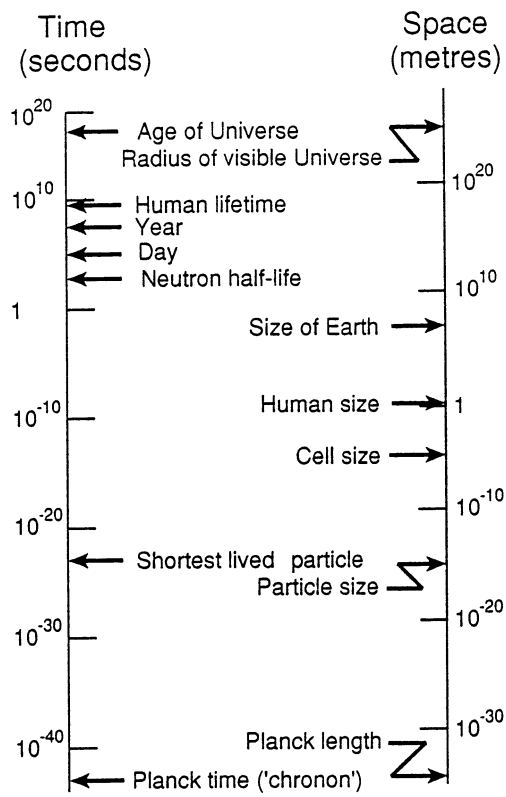



FIGURE 4. Sizes and time-scales in the Universe.

of the diagram, time-scales are shown. At the bottom of the diagram, on the left-hand side, is the very shortest time scale that is physically meaningful. This time-scale is about $10^{-43}$ of a second and is often referred to as the *Planck time-scale* or a "chronon." This time-scale is much shorter than anything experienced in particle physics. For example, the shortest-lived particles, called resonances, last for about $10^{-23}$ of a second. Further up the diagram, on the left, the day and the year are shown, and, at the top of the diagram, the present age of the Universe is shown.

On the right-hand side of the diagram, distances corresponding to these time-scales are depicted. The length corresponding to the Planck time (or chronon) is the fundamental unit of length, called the *Planck length*. These concepts of the Planck time and the Planck length fall out naturally when one tries to combine the physical theories that describe the large and the small, that is, combining Einstein's General Relativity, which describes the physics of the very large, with quantum mechanics, which describes the physics of the very small. When these theories are brought together, these Planck lengths and times turn out to be fundamental. The translation from the left-hand to the right-hand axis of the diagram is via the speed of light so that times can be translated into distances by asking how far a light signal could travel in that time.

The sizes of the physical objects represented on the diagram range from about $10^{-15}$ of a metre for the characteristic sizes of particles to about $10^{27}$ meters for the radius of the observable Universe at the present time, which is roughly the age of the Universe multiplied by the speed of light. It is intriguing to note where *we* are in the diagram, namely, the human scale. With regard to spatial dimensions, it can be seen that we are more or less in the middle of the diagram. We are enormous compared with the Planck length; even compared with the size of particles, we are very large. Yet, compared with the distance scale of the observable Universe, we are very tiny. Indeed, we are as small compared with

it as we are large compared with particles. In contrast, with regard to temporal dimensions, the human lifetime is almost as long as the Universe! People talk about the ephemeral nature of existence but, when you look at the human lifetime as shown in the diagram, it can be seen that we are not ephemeral at all — we live more or less as long as the Universe itself! Of course, this is looking on a "logarithmic scale," but this is the natural thing to do when we are concerned with such enormous ranges. To put it another way, the number of human lifetimes that make up the age of the Universe is very, very much less than the number of Planck times, or even lifetimes of the shortest-lived particles, that make up a human lifetime. Thus, we are really very stable structures in the Universe. As far as spatial sizes are concerned, we are very much in the middle — we directly experience neither the physics of the very large nor the physics of the very small. We are very much in-between. In fact, looked at logarithmically, all living objects from single cells to human beings are roughly the same in-between size.

What kinds of physics apply on these different scales? Let me introduce the diagram that summarizes the whole of physics. I have had to leave out a few details, of course, such as all the equations! But the essential basic theories that physicists use are indicated.

The key point is that, in physics, we use two very different types of procedure. To describe the small-scale behaviour, we use quantum mechanics — what I have described as the quantum level in figure 5. I shall say much more about this in the second lecture. One of the things that people say about quantum mechanics is that it is fuzzy and indeterministic, but this is not true. So long as you remain at this level, quantum theory is deterministic and precise. In its most familiar form, quantum mechanics involves use of the equation known as Schrödinger's equation, which governs the behaviour of the physical state of a quantum system — called its *quantum state* — and this is a deterministic equation. I have used the letter $U$ to describe this quantum level activity. Indeterminacy

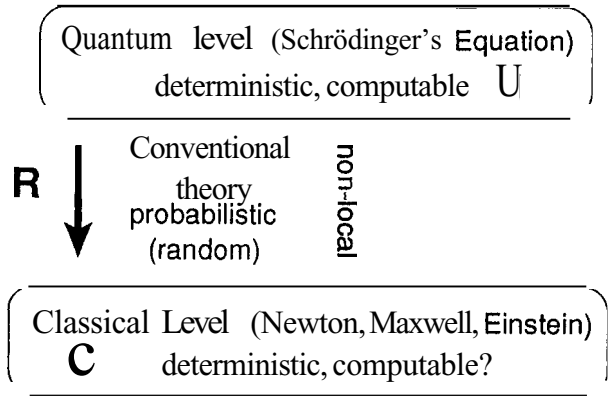



FIGURE 5.

in quantum mechanics arises only when you perform what is called "making a measurement" and that involves magnifying an event from the quantum level to the classical level. I shall say quite a lot about this in the second lecture.

On the large scale, we use classical physics, which is entirely deterministic —these classical laws include Newton's laws of motion, Maxwell's laws for the electromagnetic field, which incorporate electricity, magnetism, and light, and Einstein's theories of relativity, the Special Theory, which deals with large velocities, and the General Theory, which deals with large gravitational fields. These laws apply very, very accurately on the large scale.

Just as a footnote to figure 5, it can be seen that I have included a remark about "computability" in quantum and classical physics. This has no relevance to the present lecture or the next, but it will have importance for the third, and I shall return to the issue of computability in that lecture.

For the rest of the present lecture, I shall be primarily concerned with Einstein's theory of relativity—specifically, how theory works, its extraordinary accuracy, and something about its elegance as a physical theory. But let us first consider Newtonian theory. Newtonian physics, just as in the case of relativity, allows a space-

time description to be used. This was first precisely formulated by E.-J. Cartan for Newtonian gravity, some time after Einstein had presented his General Theory of relativity. The physics of Galileo and Newton is represented in space-time for which there is a global time coordinate, here depicted as running up the diagram (figure 6); and for each constant value of the time, there is a space section which is a Euclidean 3-space, here depicted as horizontal planes. An essential feature of the Newtonian space-time picture is that these space-slices, across the diagram, represent moments of simultaneity.

Thus, everything that occurs on Monday at noon lies on one horizontal slice through the space-time diagram; everything that happens on Tuesday at noon lies on the next slice shown in the

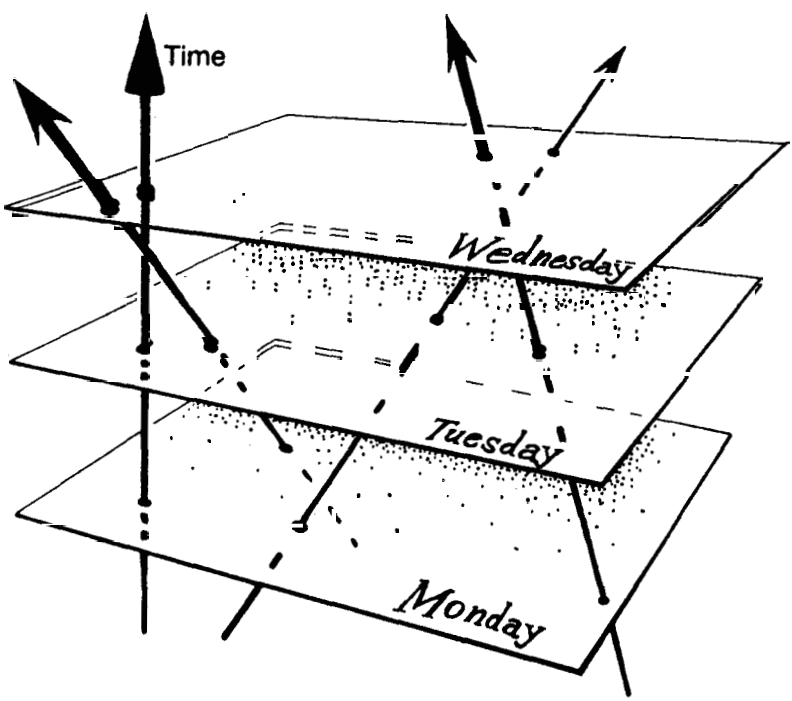



FIGURE 6. Galilean space-time: particles in uniform motion are depicted as straight lines.

diagram; and so on. Time cuts across the space-time diagram and the Euclidean sections follow one after the other as time progresses. All the observers can agree about the time when events take place, no matter how they move through the space-time, because everyone uses the same time slices to measure how time passes.

In Einstein's Special Theory of relativity, one has to adopt a different picture. In it, the space-time picture is absolutely essential —the key difference is that time is not the universal thing it is in Newtonian theory. To appreciate how the theories differ, it is necessary to understand an essential part of relativity theory, namely, those structures known as *light-cones.*

What is a light-cone? A light-cone is drawn in figure 7. We imagine a flash of light taking place at some point at some instant — that is, at an *event* in space-time — and the light waves travel outwards from this event, the source of the flash, at the speed of light. In a purely spatial picture (right-hand picture of figure 7), we can represent the paths of the light waves through space as a sphere expanding at the speed of light. We can now translate this motion of the light waves into a space-time diagram (left-hand picture of figure 7) in which time runs up the diagram and the space coordinates refer to horizontal displacements, just as

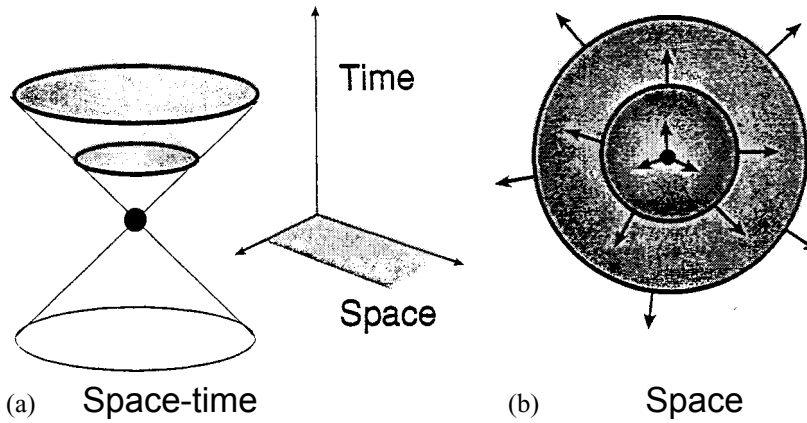



FIGURE 7. The representation of the history of a light flash in terms of its propagation in (a) space-time and (b) space.

in the Newtonian situation of figure 6. Unfortunately, in the full space-time picture, on the left of figure 7, we can represent only two spatial dimensions horizontally on the diagram, because the space of our picture is only three-dimensional, Now, we see that the flash is represented by a point (event) at the origin and that the subsequent paths of the light rays (waves) cut the horizontal "space" planes in circles, the radii of which increase at the speed of light up the diagram. It can be seen that the paths of the light rays form cones in the space-time diagram. The light-cone thus represents the history of this flash of light —light propagates away from the origin along the light-cone, which means at the speed of light, into the future. Light rays can also arrive at the origin along the light-cone from the past —that part of the light-cone is known as the past light-cone and all information carried to the observer by light waves arrives at the origin along this cone.

Light-cones represent the most important structures in space-time. In particular, they represent the limits of causal influence. The history of a particle in space-time is represented by a line travelling up the space-time diagram, and this line has to lie within the light-cone (figure 8). This is just another way of saying that a

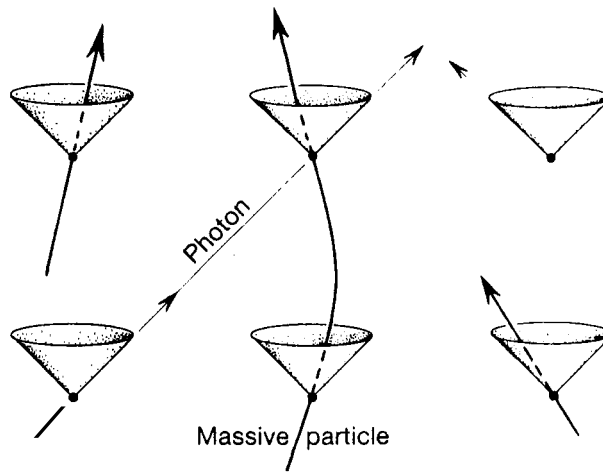



FIGURE 8. A picture of Minkowski geometry.

material particle cannot travel faster than the speed of light. No signal can travel from inside to outside the future light-cone, so that the light-cone does indeed represent the limits of causality.

There are some remarkable geometrical properties that relate to the light-cones. Let us consider two observers moving at different speeds through space-time. Unlike the case of Newtonian theory, in which the planes of simultaneity are the same for all observers, there is no absolute simultaneity in relativity. Observers moving at different speeds draw their own planes of simultaneity as different sections through space-time, as illustrated in figure 9. There is a very well-defined way of transforming from one plane to another through what is known as a *Lorentz transformation,*

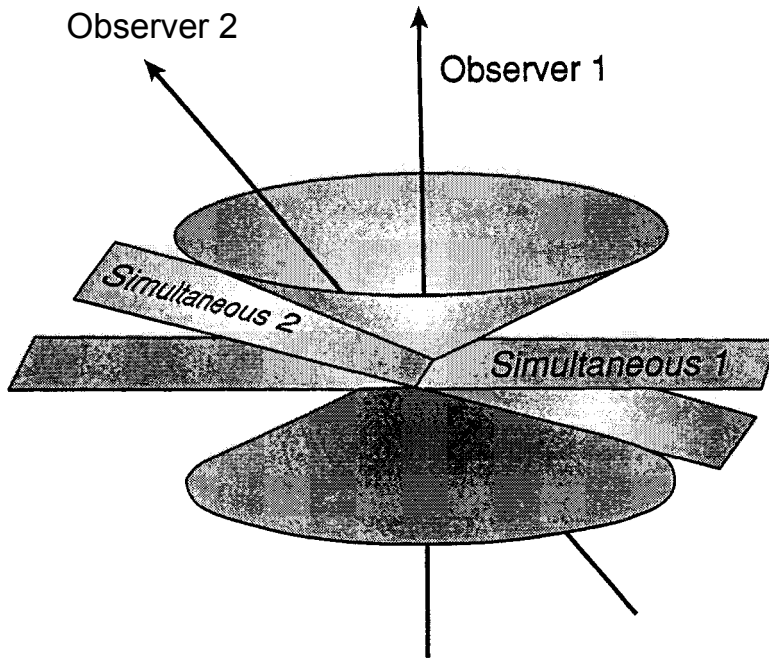



FIGURE 9. Illustrating the relativity of simultaneity according to Einstein's Special Theory of relativity, Observers 1 and 2 are moving relative to one another through space-time. Events that are simultaneous for Observer 1 are not simultaneous for Observer 2 and vice versa.

these transformations constituting what is called the *Lorentz group.* The discovery of this group was an essential feature in the discovery of Einstein's Special Theory of relativity. The Lorentz group can be understood as a group of (linear) space-time transformations, leaving a light-cone invariant.

We can also appreciate the Lorentz group from a slightly different viewpoint. As I have emphasised, the light-cones are the fundamental structures of space-time. Imagine that you are an observer located somewhere in space, looking out at the Universe. What you see are the light rays coming from the stars to your eyes. According to the space-time viewpoint, the events you observe are the intersections of the world-lines of the stars with your past light-cone, as illustrated in figure 10(a). You observe along your past light-cone the positions of the stars at particular points. These points seem to be situated on the celestial sphere that appears to surround you. Now, imagine another observer, moving at some great speed relative to you, who passes closely by you at the moment you both look out at the sky. This second observer perceives the same stars as you do, but finds them to be located in different positions on the celestial sphere —this is the effect known as *aberration.* There is a set of transformations that enables us to work out the relationship between what each of these observers sees on his or her celestial sphere. Each of these transformations is one that takes a sphere to a sphere. But it is one of a very special kind. It takes exact circles to exact circles and it preserves angles. Thus, if a pattern in the sky appears to be circular to you, then it must also appear circular to the other observer.

There is a very beautiful way of describing how this works and I illustrate it to show that there is a particular elegance in the mathematics that often underlies physics at its most fundamental level. Figure 10(c) shows a sphere with a plane drawn through its equator. We can draw figures on the surface of the sphere and then examine how they are projected to the equatorial plane from the south pole, as illustrated. This type of projection is known as

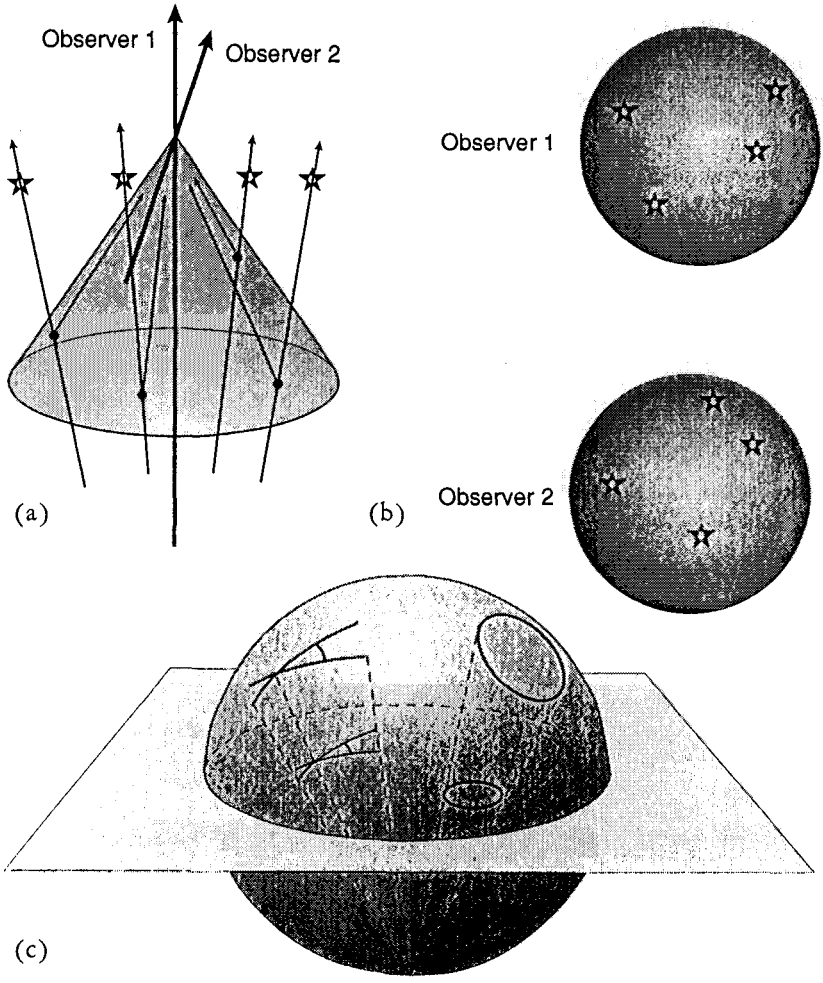



FIGURE 10. Illustrating how observations are made of the sky by Observers 1 and 2. (a) Observer 1 observes stars along the past light-cone. The points at which stars cross the light-cone are indicated by black dots. Light signals propagate from the stars to the observers along the light-cone as illustrated. Observer 2 is moving through space-time at a certain speed relative to Observer 1. (b) Illustrating the location of stars on the sky as observed by Observer 1 and Observer 2, when they are coincident at some point in space-time, (c) Illustrating the use of the Riemann sphere to locate the positions of stars on the celestial sphere for observers moving at a constant speed relative to each other.

a stereographic projection and it has some rather extraordinary properties. Circles on the sphere are projected into exact circles on the plane, and the angles between curves on the sphere are projected into exactly the same angles on the plane. As I shall discuss more fully in the second lecture, this projection allows us to label the points of the sphere by complex numbers (numbers involving the square root of $-1$) that can be used to label the points of the equatorial plane, together with "infinity," to give it the structure known as the "Riemann sphere."

For those who are interested, the transformation is

$$\mu \rightarrow \mu' = \frac{\alpha\mu + \beta}{\lambda\mu + \delta}$$

It is a fact well known to mathematicians that this transformation sends circles into circles and preserves angles. Transformations of this kind are known as Möbius transformations. For our present purposes, we need merely note the simple elegance of the form of the Lorentz (aberration) formula when written in terms of such a complex parameter $\mu$

A striking point about this way of looking at these transformations is that, according to Special Relativity, the formula is very simple, whereas, in expressing the corresponding aberration transformation according to Newtonian Mechanics, the formula would be much more complicated. It often turns out that, when you get down to the fundamentals and develop a more exact theory, the mathematics turns out to be simpler, even if the formalism appears to be more complicated in the first instance. This important point is exemplified by the contrast between Galilean and Einstein's relativity.

Thus, in the Special Theory of relativity, we have a theory that is, in many ways, simpler than Newtonian mechanics. From the point of view of mathematics, and particularly from the point of

view of group theory, it is a much nicer structure. In Special Relativity, the space-time is flat and all the light cones are lined up regularly, as illustrated in figure 8. If we now go one step further to Einstein's General Relativity, that is, the theory of space-time in the presence of gravity, the picture seems at first sight rather muddied up —the light-cones are all over the place (figure 11). Now, I have been saying that, as we develop deeper and deeper theories, the mathematics becomes simpler, but look what has happened here —I had a nice elegant piece of mathematics that has become horribly complicated. Well, that sort of thing happens —you will have to bear with me for a little while until the simplicity reappears.

Let me remind you of the fundamental ingredients of Einstein's theory of gravity. One basic ingredient is called Galileo's Principle of Equivalence. In figure 12, I show Galileo leaning over from the top of the Tower of Pisa dropping large and small rocks. Whether or not he actually performed this experiment, he certainly well understood that, if the effects of air resistance are ignored, the two rocks would fall to the ground in the same time. If you happened to be sitting on one of these rocks looking at the other one as they fall together, you would observe the other rock hovering in front of you (I have shown a camcorder attached to one of the rocks to make the observation). Nowadays, with space travel, this is a very familiar phenomenon —just recently, we have seen a British-born astronaut walking in space, and, just like the big rock and the little rock, the spaceship hovers in front of the astronaut —this is exactly the same phenomenon as Galileo's Principle of Equivalence.

Thus, if you look at gravity in the right way, that is, in a falling frame of reference, it seems to disappear right in front of your eyes. This is indeed correct. But Einstein's theory does *not* tell you that gravity disappears — it only tells you that the *force* of gravity disappears. There is something left and that is the tidal effect of gravity.

Let me introduce a little bit more mathematics, but not much. We need to describe the curvature of space-time and this is de-

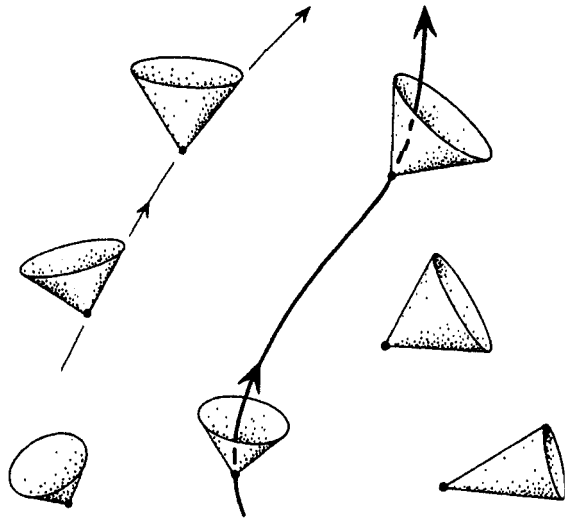


FIGURE 11. A picture of curved space-time.

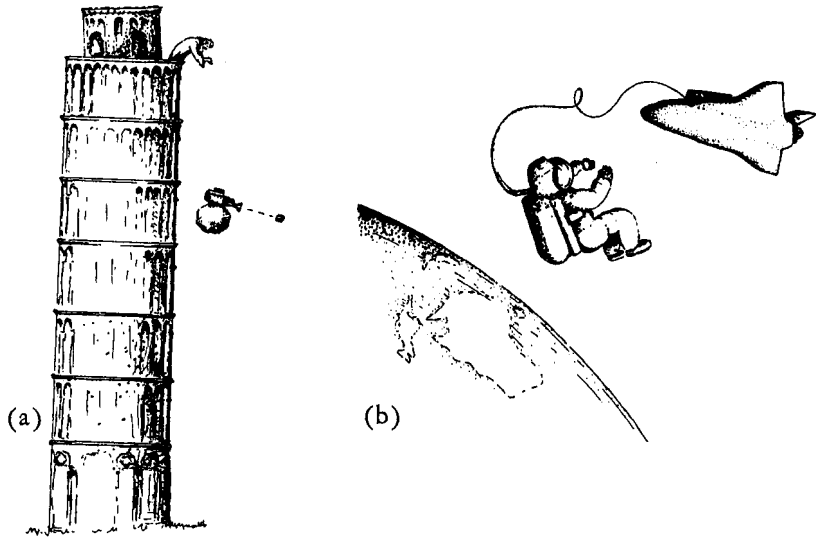



FIGURE 12. (a) Galileo dropping two rocks (and a camcorder) from the Leaning Tower of Pisa. (b) The astronaut sees the space-vehicle hover before him or her, seemingly unaffected by gravity.

scribed by an object known as a tensor which I have called *Riemann* in the following equation. It is actually called the Riemann curvature tensor but I will not tell you what it is except that it is represented by a capital $R$ with a number of indices stuck on the bottom, which are indicated by the dots. The Riemann curvature tensor is made up of two pieces. One of the pieces is called the *Weyl* curvature and the other piece is called the *Ricci* curvature, and we have the (schematic) equation

$$\begin{aligned} \textit{Riemann} &= \textit{Weyl} + \textit{Ricci} \\ R\ldots &= C\ \ldots\ + R..g.. \end{aligned}$$

Formally, C. . . . and $R$. . are the Weyl and Ricci curvature tensors, respectively, and g. . is the metric tensor.

The Weyl curvature measures the tidal effect. What is the "tidal" effect? Recall that, from the astronaut's point of view, it seems that gravity has been abolished, but that is not quite true. Imagine that the astronaut is surrounded by a sphere of particles, which are initially at rest with respect to the astronaut. Now, at first they will just hover there but soon they will start to accelerate because of the slight differences in the gravitational attraction of the Earth at different points in the sphere. (Notice that I am describing the effect in Newtonian language, but that is quite adequate.) These slight differences cause the original sphere of particles to become distorted into an elliptical arrangement, as illustrated in figure 13(a) .

This distortion occurs partly because of the slightly greater attraction of the Earth for those particles closer to the Earth and lesser attraction for those further away, and partly because, at the sides of the sphere, the Earth's attraction acts slightly inwards. This causes the sphere to be distorted into an ellipsoid. It is called the tidal effect for the very good reason that, if you replace the Earth by the Moon and the sphere of particles by the Earth with its oceans, then the effect of the Moon on the surface of the oceans is

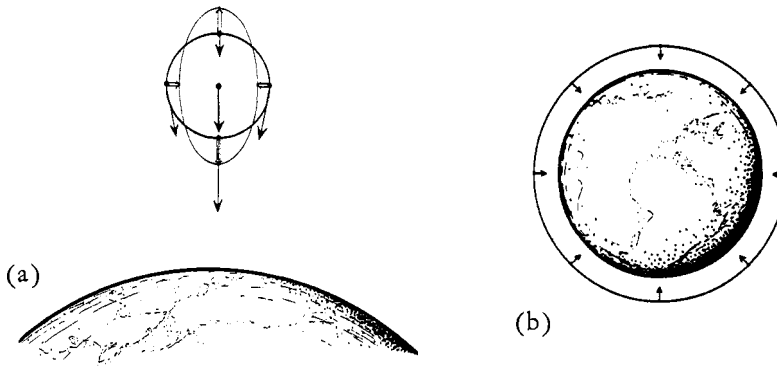



FIGURE 13. (a) The tidal effect. Double arrows show relative acceleration. (b) When the sphere surrounds matter (here on Earth), there is a net inward acceleration.

exactly the gravitational effect upon the sphere of particles —the sea surface closest to the Moon is pulled towards it, whereas the ocean surface on the other side of the Earth is, in effect, pushed away from it. The effect causes the sea surface to bulge out on either side of the Earth and is the cause of the two high tides that occur each day.

The effects of gravity, from Einstein's point of view, are simply this tidal effect. It is defined precisely by the Weyl curvature, that is, the part $C_{\ldots}$ of the Riemann curvature. This part of the curvature tensor is volume-preserving —that is, if you work out the initial accelerations of the particles of the sphere, the volume of the sphere and the volume of the ellipsoid into which it is distorted are initially the same.

The remaining part of the curvature is known as the *Ricci* curvature and it has a volume-reducing effect. From figure 13(b), it can be seen that if, instead of being at the bottom of the diagram, the Earth were inside the sphere of particles, the volume of the sphere of particles would be reduced as the particles accelerate inwards. The amount of this reduction in volume is a measure of the Ricci curvature. Einstein's theory tells us that the Ricci curva-

ture is determined by the amount of matter present within a small sphere about that point in space. In other words, the density of matter, appropriately defined, tells us how the particles are accelerated inwards at that point in space. Einstein's theory is almost the same as Newton's when expressed in this way.

This is how Einstein formulates his theory of gravity — it is expressed in terms of the tidal effects, which are measurements of the local space-time curvature. It is crucial that we have to think in terms of the curvature of four-dimensional space-time. This was shown schematically in figure 11 — we think of the lines that represent the world lines of particles and the ways in which these paths are distorted as a measurement of the curvature of space-time. Thus, Einstein's theory is essentially a geometric theory of four-dimensional space-time — it is an extraordinarily beautiful theory mathematically.

The history of Einstein's discovery of the theory of General Relativity contains an important moral. It was first fully formulated in 1915. It was not motivated by any observational need but by various aesthetic, geometric, and physical desiderata. The key ingredients were Galileo's Principle of Equivalence, exemplified by his dropping rocks of different masses (figure 12), and the ideas of non-Euclidean geometry, which is the natural language for describing the curvature of space-time. There was not a great deal on the observational side in 1915. Once General Relativity was formulated in its final form, it was realised that there were the three key observational tests of the theory. The perihelion of the orbit of Mercury is swung around, or precesses, in a way that could not be explained by the Newtonian gravitational influence of the other planets — General Relativity predicts exactly the observed precession. The paths of light rays are bent by the Sun and this was the reason for the famous eclipse expedition of 1919, led by Arthur Eddington, which found a result consistent with Einstein's prediction (figure 14). The third test was the prediction that clocks run slow in a gravitational potential — that is, a clock closer to the

ground runs slow with respect to a clock at the top of a tower. This effect has also been measured experimentally. These were never, however, very impressive tests —the effects were always very small and various different theories can give the same results.

The situation has now changed dramatically —in 1993, Russell Hulse and Joseph Taylor were awarded the Nobel prize for a most

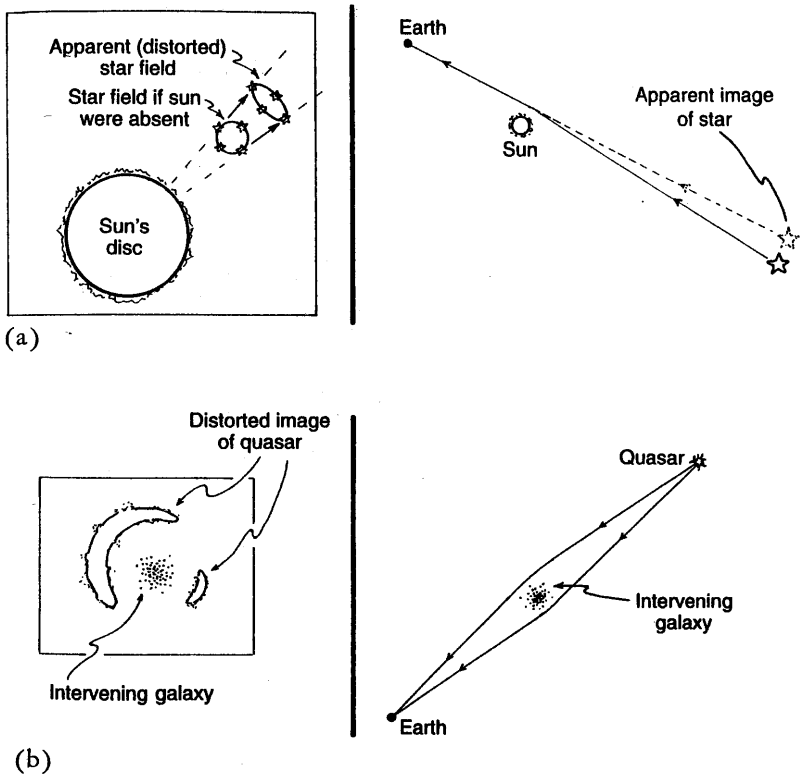
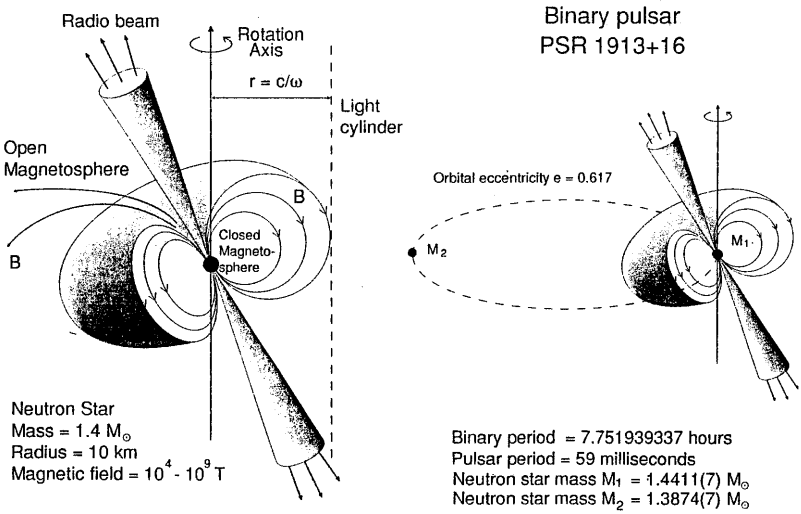



FIGURE 14. (a) A direct observational effect of light-cone tilting. The *Weyl* space-time curvature manifests itself as a distortion of the distant star field, here owing to the light-bending effect of the Sun's gravitational field. A circular pattern of stars would get distorted into an elliptical one. (b) Einstein's light-bending effect is now an important tool in observational astronomy. The mass of the intervening galaxy may be estimated by how much it distorts the image of a distant quasar.

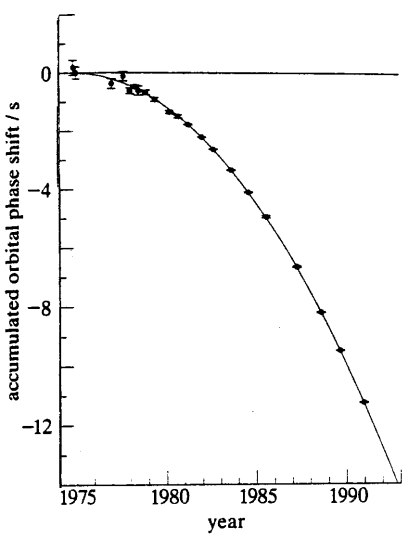
extraordinary series of observations. Figure 15 shows the binary pulsar known as PSR 1913+16 —it consists of a pair of neutron stars, each of which is an enormously dense star that has mass about that of the Sun but is only a few kilometres in diameter. The neutron stars orbit around their common centre of gravity in highly elliptical orbits. One of them has a very strong magnetic field and particles get swung round and emit intense radiation that travels to the Earth, some 30,000 light years away, where it is observed as a series of well-defined pulses. All sorts of very precise observations have been made of the arrival times of these pulses. In particular, all the properties of the orbits of the two neutron stars can be worked out as well as all the tiny corrections due to General Relativity.

There is, in addition, a feature that is completely unique to General relativity, and not present at all in the Newtonian theory of gravity. That is that objects in orbit about each other radiate away energy in the form of gravitational waves. These are like light waves but are ripples in space-time rather than ripples in the electromagnetic field. These waves take energy away from the system at a rate that can be precisely calculated according to Einstein's theory, and the rate of loss of energy of the binary neutron star system agrees very precisely with the observations, as illustrated by figure 15(c), which shows the speed-up of the orbital period of the neutron stars, measured over twenty years of observation. These signals can be timed so precisely that, over twenty years, the accuracy with which the theory is known to be correct amounts to about one part in $10^{14}$ This makes General Relativity the most accurately tested theory known to science.

There is a moral in this story —Einstein's motivations for devoting eight or more years of his life to deriving the General Theory were not observational or experimental. Sometimes people argue, "Well, physicists look for patterns in their experimental results and then they find some nice theory that agrees with these. Maybe this explains why mathematics and physics work so well




(a) (b)




(c)

FIGURE 15. (a) A schematic diagram illustrating the properties of those neutron stars that are radio pulsars. Radio emission is emitted along the poles of the magnetic dipole that is misaligned with respect to the rotation axis of the neutron star. Sharply defined pulses are observed when the narrow beam of radiation is swept across the line of sight to the observer. (b) A schematic representation of the binary pulsar PSR 1913 + 16. The properties of the two neutron stars have been derived from very precise timing of the arrival times of the pulses using effects that are only present in Einstein's General Relativity. (c) The change of phase of the arrival times of the pulses from the binary pulsar PSR 1913 + 16, compared with the expected change due to the emission of gravitational radiation by the binary neutron star system.

together." But, in this case, things were not like that at all. The theory was developed originally without any observational motivation —the mathematical theory is very elegant and it is physically very well motivated. The point is that the mathematical structure is just there in nature, the theory really is out there in space—it has not been imposed upon nature by anyone. That is one of the essential points of this lecture. Einstein revealed something that was there. Moreover, it was not just some minor piece of physics he discovered —it is the most fundamental thing that we have in nature, the nature of space and time.

Here is a very clear case —it goes back to my original diagram concerning the relation between the world of mathematics and the physical world (figure 3). In General Relativity, we have some kind of structure that really does underlie the behaviour of the physical world in an extraordinarily precise way. The way in which these fundamental features of our world are discovered is often not by looking at the way in which nature behaves, although that is obviously very important. One has to be prepared to throw out theories that might appeal for all sorts of other reasons but that do not fit the facts. But here we have a theory that does fit the facts with extraordinary accuracy. The accuracy involved is about twice as many figures as one has in Newtonian theory; in other words, General Relativity is known to be correct to one part in $10^{14}$ whereas Newtonian theory was tested only to one part in $10^{7}$. The improvement is similar to the increase in accuracy with which Newton's theory was known to be correct between the seventeenth century and now. Newton knew his theory was correct to about one part in 1,000, whereas now it is known to be accurate to one part in $10^{7}$.

Einstein's General Relativity is just a theory, of course. What about the structure of the actual world? I said this lecture would not be botanical but, if I talk about the Universe as a whole, that is not being botanical, since I will consider only the one Universe as a whole that is given to us. There are three types of standard model

that come out of Einstein's theory and these are defined by one parameter, which is, in effect, the one denoted by $k$ in figure 16. There is another parameter that sometimes appears in cosmological arguments that is known as the cosmological constant. Einstein regarded his introduction of the cosmological constant into his equations of General Relativity as his greatest mistake and so I shall leave it out too. If we are forced to bring it back, well, we shall have to live with it.

Assuming the cosmological constant is zero, the three types of Universe, which are described by the constant $k$, are illustrated in figure 16. In the diagram, $k$ takes values 1, 0, and -1, because all the other properties of the models have been scaled away. A better way would have been to talk about the age or scale of the Universe, and then one would have a continuous parameter but, qualitatively, the three different models can be thought of as being defined by the curvature of the space sections of the Universe. If the space sections of Universe are flat, they have zero curvature and $k = 0$. If the space sections are positively curved, meaning that the Universe closes in on itself, then $k = +1$. In this case the Universe has an initial singular state, the Big Bang, which marks the beginning of the Universe. It expands to a maximum size and then recollapses to a Big Crunch. Alternatively, there is the $k = -1$ case, in which the Universe expands forever. The $k = 0$ case is the limiting boundary between the $k = 1$ and $k = -1$ cases. Beside the diagrams, I have shown the radius-time relations for these three types of Universe in figure 16(d). The radius can be thought of as some typical scale in the Universe and it can be seen that only the case $k = +1$ collapses to a Big Crunch, while the other two expand indefinitely.

I want to consider the $k = -1$ case in a little more detail — it is perhaps the most difficult of the three to come to terms with. There are two reasons for being interested in this case particularly. One reason is that, if you take the observations as they exist at the moment at their face value, it is the preferred model. According

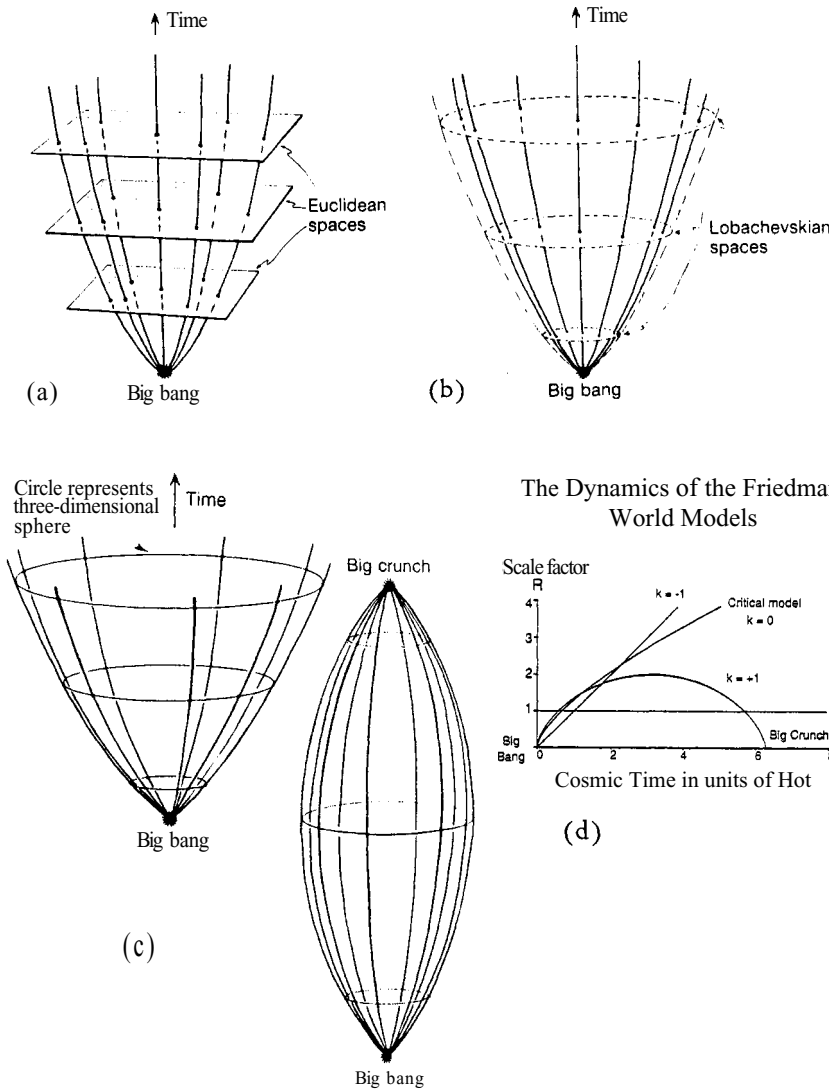



FIGURE 16. (a) Space-time picture of an expancaLig universe with Euclidean spatial sections (two space dimensions depicted), $k = 0$. (b) As in (a) but for an expanding universe with Lobachevskian spatial sections, $k = -1$. (c) As in (a), but for an expanding universe with spherical spatial sections $k = +1$. (d) The dynamics of the three different types of Friedman model.

to General Relativity, the curvature of space is determined by the amount of matter present in the Universe and there doesn't seem to be enough to close the geometry of the Universe. Now, it may be that there is a lot of dark or hidden matter, which we do not yet know about. In this case, the Universe could be one of the other models but, if there is not a lot of extra matter, much more than we believe must be present within the optical images of galaxies, then the Universe would have $k = -1$. The other reason is that it is the one I like the best! The properties of $k = -1$ geometries are particularly elegant.

What do the $k = -1$ universes look like? Their spatial sections have what is known as hyperbolic or Lobachevski geometry. To get a picture of a Lobachevski geometry it is best to look at one of Escher's prints. He made a number of prints that he called *Circle Limits,* and *Circle Limit 4* is shown in figure 17. This is Escher's description of the Universe-you see it is full of angels and devils!

A point to note is that it looks as though the picture gets very crowded towards the edge of the limit circle. This occurs because this representation of hyperbolic space is drawn on an ordinary plane sheet of paper, in other words, in Euclidean space. What you have to imagine is that all the devils are supposed to be actually exactly the same size and shape so that, if you happened to live in this Universe towards the edge of the diagram, they would look exactly the same to you as the ones in the middle of the diagram. This picture gives some impression of what is going on in Lobachevski geometry —as you walk from the centre out to the edge, you have to imagine that, because of the way the picture of the geometry has had to be distorted, the actual geometry there is exactly the same as it is in the middle, so that the geometry all about you remains the same no matter how you move.

This is perhaps the most surprising example of a well-defined geometry. But Euclidean geometry is, in its way, just as remarkable. Euclidean geometry provides a wonderful illustration of the

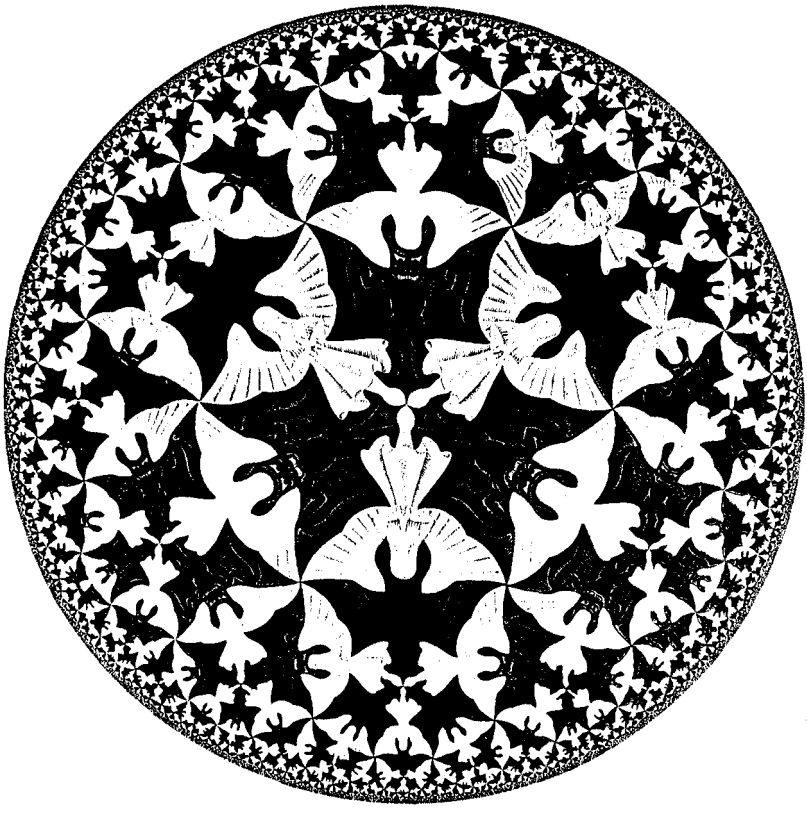


FIGURE 17. *Circle Limit 4* by M. C. Escher.

relation between mathematics and physics. The geometry is a part of mathematics, but the Greeks also thought of it as a description of the way the world is. Indeed, it turns out to be an extraordinarily accurate description of the way the world actually is—not utterly accurate because Einstein's theory tells us that space-time is slightly curved in various ways, but it is an extraordinarily accurate description of the world nevertheless. People used to worry about whether or not other geometries were possible. In particular, they worried about what is known as *Euclid's fifth postulate.* This can be reformulated as the statement that, if there is a

line in a plane and there is a point outside that line, then there is a unique parallel to this line through that point. People used to think that maybe this could be proved from the other more obvious axioms of Euclidean geometry. It turns out that it is not possible, and from this the notion of non-Euclidean geometry arose.

In non-Euclidean geometries, the angles of a triangle do not add up to 180°. This is another example where you think that things must become more complicated because, in Euclidean geometry, the angles of a triangle do add up to 180°. But then, in the non-Euclidean geometry, if you take the sum of the angles of a triangle away from 180° you find that this difference is proportional to the area of the triangle. In Euclidean geometry, the area of a triangle is a complicated thing to write down in terms of angles and lengths. In non-Euclidean, Lobachevskian, geometry, there is this wonderfully simple formula, due to Johann Heinrich Lambert, which enables the area of the triangle to be found. In fact, Lambert derived his formula before non-Euclidean geometry was discovered and I have never quite understood that!

There is another very important point here, which concerns the real numbers. These are absolutely fundamental to Euclidean geometry. They were essentially introduced by Eudoxus in the fourth century B.C. and they are still with us. They are the numbers that describe all our physics. As we shall see later, complex numbers are needed too, but they are based upon real numbers.

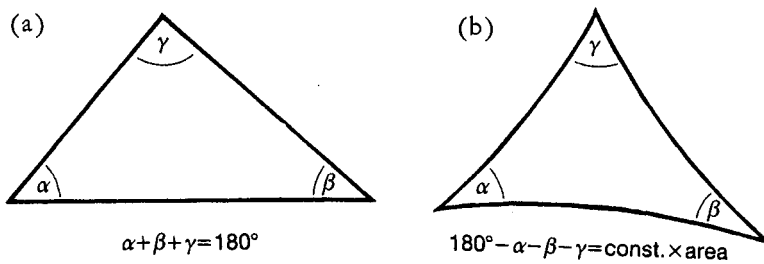



FIGURE 18. (a) A triangle in Euclidean space. (b) A triangle in Lobachevskian space.

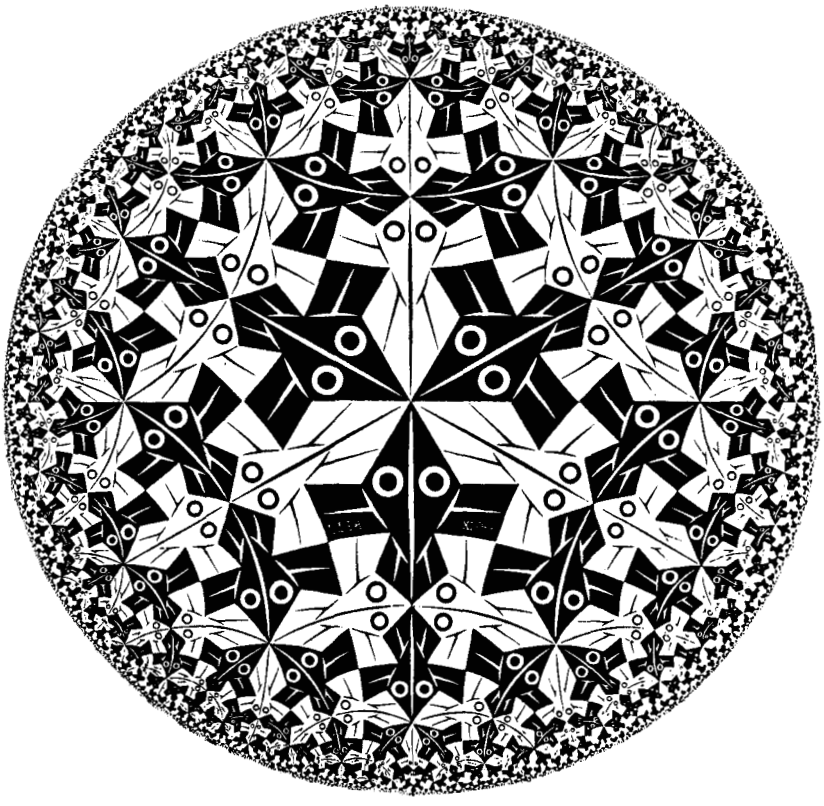


FIGURE 19. *Circle Limit 1* by M. C. Escher.

Let us look at another of the Escher prints to see how the Lobachevski geometry works. Figure 19 is even nicer than figure 18 for understanding this geometry because the "straight lines" are more obvious. They are represented by arcs of circles that meet the boundary at right angles, So, if you were a Lobachevskian person, and lived in this geometry, what you would think of as a straight line would be one of these arcs. You can see these clearly in figure 19 —some of them are straight lines through the centre of the diagram, but all the others are curved arcs. Some of these "straight lines" are shown in figure 20. In that diagram, I have

indicated a point that is not on the straight line crossing the diagram. Lobachevskian people can draw two (and more) separate lines parallel to the diagonal through that point, as I have indicated. Thus, the parallel postulate is violated in this geometry. Furthermore, you can draw triangles and work out the sums of angles of the triangles in order to work out their areas. This may give you some taste for the nature of hyperbolic geometry.

Let me give another example. I said that I like hyperbolic, Lobachevskian geometry the best. One of the reasons is that its group of symmetries is exactly the same as the one that we have already encountered, namely, the Lorentz group —the group of Special Relativity, or the symmetry group of the light-cones of relativity. To see that it is, I have drawn a light-cone in figure 21 but with some extra bits drawn on. I have had to suppress one of the space dimensions in order to depict it in three-dimensional space.

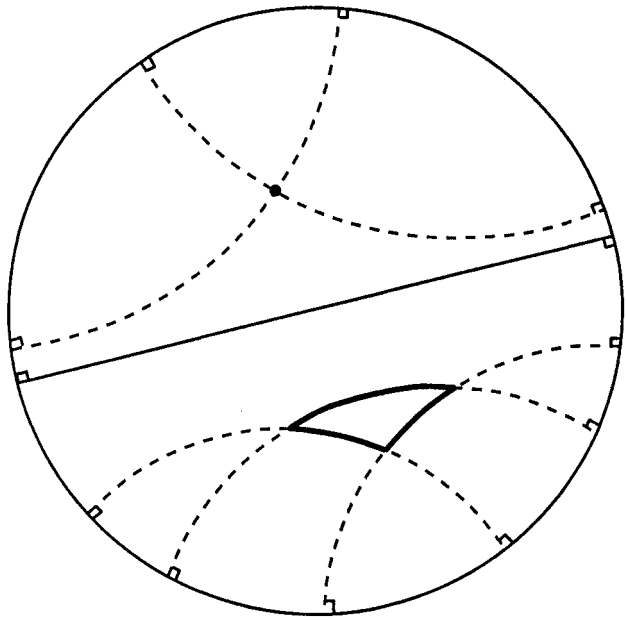


FIGURE 20. Illustrating the geometry of Lobachevskian (hyperbolic) space.

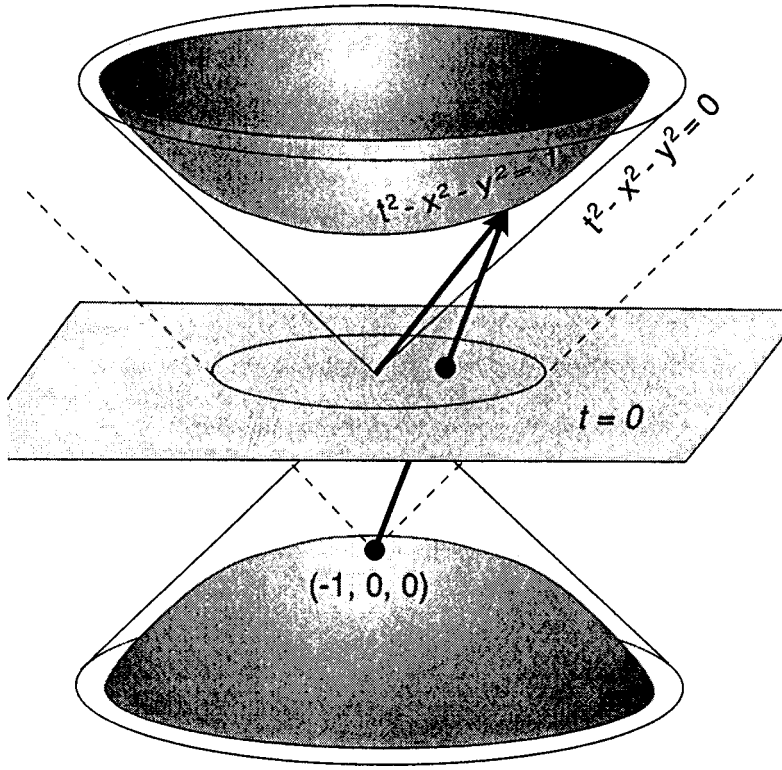



FIGURE 21. Illustrating Lobachevskian space-time geometry. The circle drawn on the plane $t=0$ is the Poincaré disc.

The light-cone is described by the usual equation shown on the diagram:

$$t^2 - x^2 - y^2 = 0$$

The bowl-shaped surfaces shown above and below are located at "unit distance" from the origin in this Minkowskian geometry. ("Distance" in Minkowskian geometry is actually *time*—the proper time that is physically measured by moving clocks.) Thus, these surfaces represent the surface of a "sphere" for the Minkowskian geometry. It turns out that the intrinsic geometry of the

"sphere" is actually Lobachevskian (hyperbolic) geometry. If you consider an ordinary sphere in Euclidean space, you can rotate it around, and the group of symmetries is that of rotating the sphere around. In the geometry of figure 21, the group of symmetries is the group of symmetries associated with the surface shown in the diagram, in other words, with the Lorentz group of rotations. This symmetry group describes how space and time transform when a particular point in space-time is fixed —rotating the space-time about in different ways. We now see, with this representation, that the group of symmetries of Lobachevskian space is essentially just the same as the Lorentz group.

We can carry out a Minkowskian version of the stereographic projection, as illustrated in figure 10(b), The equivalent of the south pole is now the point at $(-1, 0, 0)$ and we project points from the upper bowl-shaped surface to the flat surface at $t = 0$, which is the equivalent of the equatorial plane in figure 10(b). In this procedure, we project all the points on the upper surface to the plane at $t = 0$. The projected points all lie inside a disc in the plane at $t = 0$, and this disc is sometimes referred to as the Poincaré disc. This is precisely how Escher's circle limit diagrams come about —the entire hyperbolic (Lobachevskian) surface has been mapped onto the Poincaré disc. Furthermore, this mapping does all the things that the projection of figure 10(b) does —it preserves angles and circles and it all comes out geometrically in a very nice way. Well, perhaps I am getting carried away here by my enthusiasm — I am afraid that is what mathematicians do when they get stuck into something.

The intriguing point is that, when you get carried away by something like the geometry of the above problem, the analysis and the results have an elegance that sustains them, while analyses that do not possess this mathematical elegance peter out. There is something particularly elegant about hyperbolic geometry. It would be awfully nice, at least to the likes of me, if the Universe were built that way too. Let me say that I have various other rea-

sons for believing this. Many other people do not like these open, hyperbolic universes —they frequently prefer closed universes, such as those illustrated in figure 16(c), which are nice and cosy. Well, actually, the closed universes are pretty big still. Alternatively, many people like flat world models as in figure 16(a) because there is a certain type of theory of the early Universe, the *inflationary theory,* which suggests that the geometry of the Universe should be flat. I should say that I do not really believe these theories.

These three standard types of model of the Universe are what are known as the Friedman models and they are characterised by the fact that they are very, very symmetrical. They are initially expanding models but at any moment the Universe is perfectly uniform everywhere. This assumption is built into the structure of the Friedman models and it is known as the *cosmological principle.* Wherever you are, the Friedman universe looks the same in all directions. It turns out that our actual Universe is like this to a remarkable degree. If Einstein's equations are right, and I have shown that the theory agrees with observation to a quite remarkable degree, then we are led to take the Friedman models seriously. All these models have this awkward feature, known as the Big Bang, where everything goes wrong, right at the beginning. The Universe is infinitely dense, infinitely hot, and so on —something has gone badly wrong with the theory. Nonetheless, if you accept that this very hot, dense phase took place, you can make predictions about what the thermal content of the Universe should be today and one of these expectations is that there should be a uniform background of black-body radiation all about us at the present day. Precisely this type of radiation was discovered by Arno Penzias and Robert Wilson in 1965. The most recent observations of the spectrum of this radiation, which is known as the Cosmic Microwave Background Radiation, by the COBE satellite show that it has precisely a black body spectrum with quite extraordinary precision.

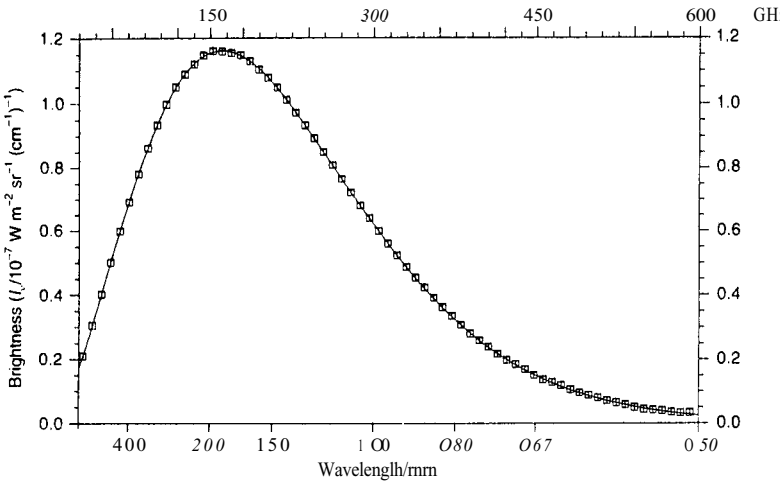



FIGURE 22. The precise agreement between the COBE measurements of the spectrum of the cosmic microwave background radiation and the expected "thermal" nature of the Big Bang's radiation.

All cosmologists interpret the existence of this radiation as evidence that our Universe went through a hot, dense phase. This radiation is thus telling us something about the nature of the early Universe —it is not telling us everything, but something like the Big Bang did take place. In other words, the Universe must have been very like the models illustrated in figure 16.

There is one other very important discovery made by the COBE satellite and this is that, although the Cosmic Microwave Background Radiation is remarkably uniform and its properties can all be accounted for very beautifully mathematically, the Universe is not quite perfectly uniform. There are tiny but real irregularities in the distribution of the radiation over the sky. In fact, we expect that these tiny irregularities must be present in the early Universe —we are here to observe the Universe and we are certainly not just a uniform smudge. The Universe is probably more like the pictures illustrated in figure 23. To show how open-minded I am, I am using as examples both an open and a closed Universe.

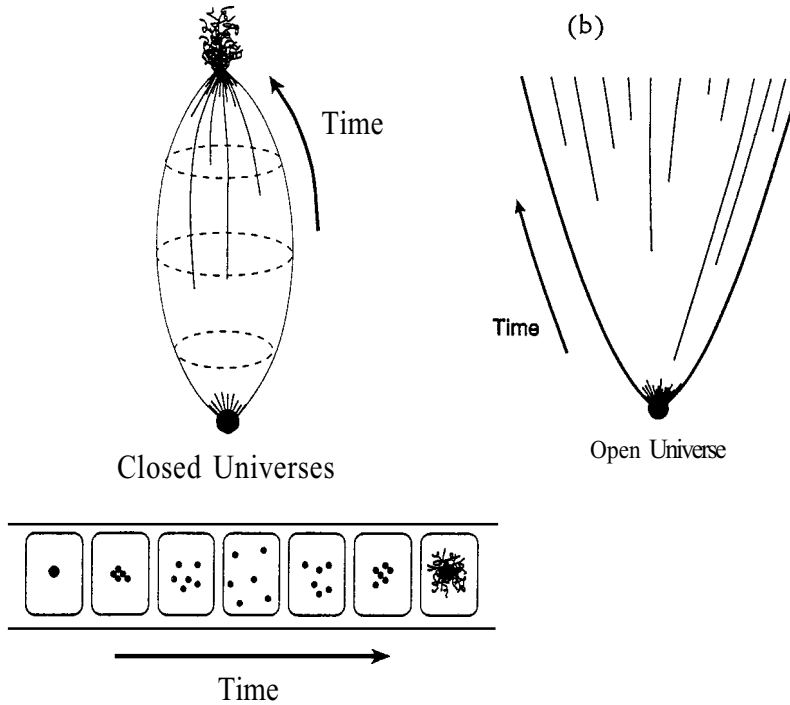



FIGURE 23. (a) The evolution of a closed world model with the formation of black holes as objects of various types reach the end points of their evolution. It can be seen that there is expected to be a horrible mess at the Big Crunch. The sequence of events is also shown as a "film-strip." (b) The evolution of the open model with the formation of black holes.

In the closed Universe, the irregularities will develop to form real observable structures —stars, galaxies, and the like —and, after a while, black holes will form, through the collapse of stars, through the accumulation of mass at the centres of galaxies, and so on. These black holes all have singular centres, much like the Big Bang in reverse. However, it is not as simple as that. According to the picture we have developed, the initial Big Bang is a nice, symmetrical, uniform state but the end point of the closed model is a horrible mess —all the black holes finally coming together and producing an incredible jumble at the final Big Crunch, as in

figure 23(a). The evolution of this closed model is illustrated schematically by the film strip shown in figure 23(b). In the case of an open universe model, the black holes are still formed as well; there is still an initial singularity and singularities formed at the centres of the black holes.

I emphasise these features of the standard Friedman models to show that there is a great difference between what we seem to see in the initial state and what we expect to find in the remote future. This problem is connected with the fundamental law of physics known as the Second Law of Thermodynamics.

We can understand this law in simple everyday terms. Imagine a glass of wine perched on the edge of the table. It might fall off the table, smash to pieces, and the wine spill all over the carpet (figure 24). There is nothing in Newtonian physics that tells us that the reverse process cannot happen. However, that is never observed —you never see wine glasses reassembling themselves and the wine being sucked up out of the carpet and into the reassembled glass. So far as the detailed laws of physics are concerned, the one direction of time is just as good as the other. To understand this difference, we need the Second Law of Thermo-

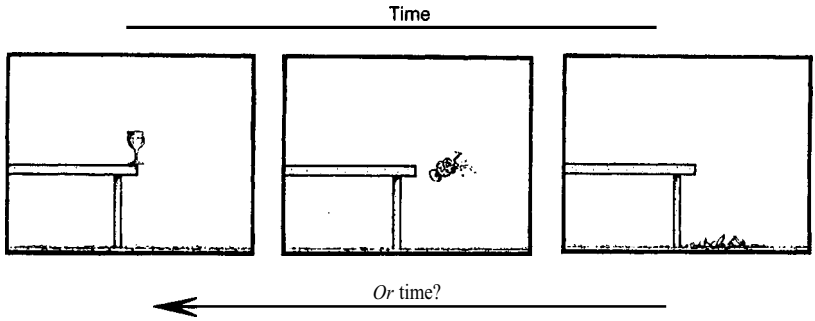



FIGURE 24. The laws of mechanics are time-reversible; yet the time-ordering of such a scene from the right frame to the left is something that is never experienced, whereas that from the left frame to the right would be commonplace.

dynamics, which tells us that the entropy of the system increases with time. This quantity called entropy is lower when the glass is on the table as compared with when it is shattered on the floor. According to the Second Law of Thermodynamics, the entropy of the system has increased. Roughly speaking, entropy is a measure of the disorder of a system. To express this concept more precisely, we have to introduce the concept of a *phase space.*

A phase space is a space of an enormous number of dimensions and each point of this multidimensional space describes the positions and momenta of all the particles that make up the system under consideration. In figure 25, we have selected a particular point in this huge phase space that represents where all the particles are located and how they are moving. As the system of particles evolves, the point moves to somewhere else in the phase space and I have shown it wiggling about from one point in phase space to another.

This wiggly line represents the ordinary evolution of the system of particles. There is no entropy there yet. To get entropy, we

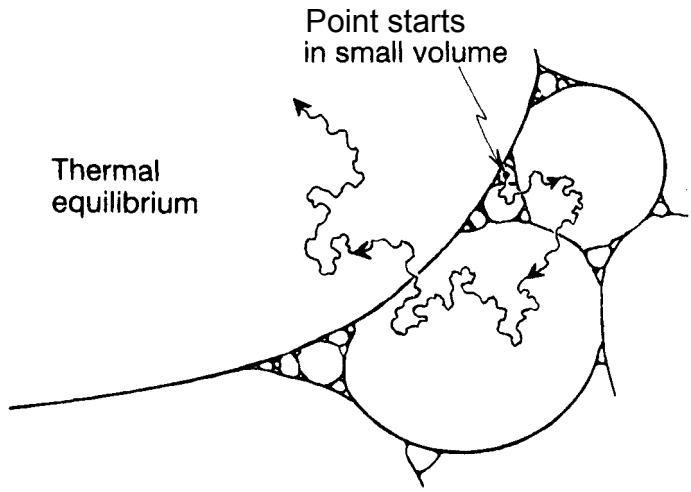



FIGURE 25. The Second Law of Thermodynamics in action: as time evolves, the phase-space point enters compartments of larger and larger volumes. Consequently, the entropy continually increases.

have to draw little bubbles around regions by lumping together those different states that you cannot tell apart. That may seem a bit obscure —what do you mean by "Cannot tell apart"? Surely that depends upon those who are looking and how carefully they look? Well, it is one of the slightly tricky questions of theoretical physics to say exactly what you mean by entropy. Essentially, what is meant is that you have to group states together according to what is known as "coarse-graining," that is, according to those things that you cannot tell apart. You take all those that, say, lie in this region here and lump them together, you look at the volume of that region, take the logarithm of the volume, and multiply it by the constant known as Boltzmann's constant, and that is the entropy. What the Second Law of Thermodynamics tells us is that the entropy increases. What it is telling you is actually something rather silly —all it is saying is that, if the system starts off in a little tiny box and it is allowed to evolve, it moves into bigger and bigger boxes. It is very likely that this happens because, if you look at the problem carefully, the bigger boxes are absolutely stupendously huger than the neighbouring little boxes. So, if you find yourself in one of the big boxes, there is absolutely no chance of getting back into a smaller box. That is all there is to it. The system just wanders about in phase space getting into bigger and bigger boxes. That is what the Second Law is telling us. Or is it?

Actually, that is only half the explanation. It tells us that, if we know the state of the system now, we can tell the most likely state in the future. But it tells us the completely wrong answer if we try to use the same argument backwards. Suppose the glass is sitting on the edge of the table. We can ask, "What is the most likely way by which it could have got there?" If you use the argument we have just given backwards, you would conclude that the most likely thing is that it started as a great mess on the carpet and then picked itself up off the carpet and reassembled itself on the table. This is clearly not the correct explanation —the correct ex-

planation is that someone put it there. And that person put it there for some reason, which was in turn due to some other reason, and so on. The chain of reasoning goes back and back to lower and lower entropy states in the past. The correct physical curve is illustrated in figure 26 —the entropy goes down and down and down in the past.

The reason why the entropy increases in the future is explained by moving into larger and larger boxes — why it goes down in the past is something completely different. There must have been something that pulled it down in the past. What pulled it down in the past? As we go into the past, the entropy gets smaller and smaller until eventually we end up at the Big Bang.

There must have been something very, very special about the Big Bang, but exactly what that was is a controversial issue. One popular theory, which I said I did not believe but which a lot of people are very keen on, is the idea of the inflationary universe. The idea is that the Universe is so uniform on the large scale because of something that was supposed to have taken place in the

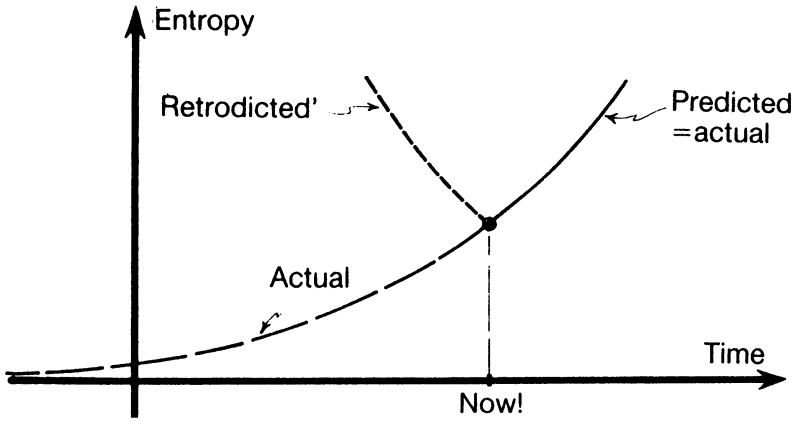



FIGURE 26. If we use the argument depicted in figure 25 in the reverse direction in time, we "retrodict" that the entropy should also increase in the *past,* from its value now. This is in gross contradiction with observation.

very earliest phases of the expansion of the Universe. It is supposed that an absolutely enormous expansion took place when the Universe was only about $10^{-36}$ seconds old and the idea is that, no matter what the Universe looked like in these very early stages, if you expand it by a huge factor of about $10^{60}$, then it will look flat. In fact, this is one reason why these people like the flat Universe.

But, as it stands, the argument does not do what it is supposed to do —what you would expect in this initial state, if it were randomly chosen, would be a horrendous mess and, if you expand that mess by this huge factor, it still remains a complete mess. In fact, it looks worse and worse the more it expands (figure 27).

So the argument by itself does not explain why the Universe is so uniform. We need a theory that tells you what the Big Bang was really like. We do not know what that theory is, but we know that it has to involve a combination of large-scale and small-scale physics. It has to involve quantum physics as well as classical physics. Furthermore, I would claim that the theory must also have as one of its implications that the Big Bang was as uniform as we observe it to be. Maybe such a theory will end up producing a hyperbolic, Lobachevskian universe, like the picture I prefer, but I shall not insist upon that.

Let us return to the pictures of the closed and open Universes again (figure 28). In addition, I have included a picture of the formation of a black hole, which will be well known to the ex-

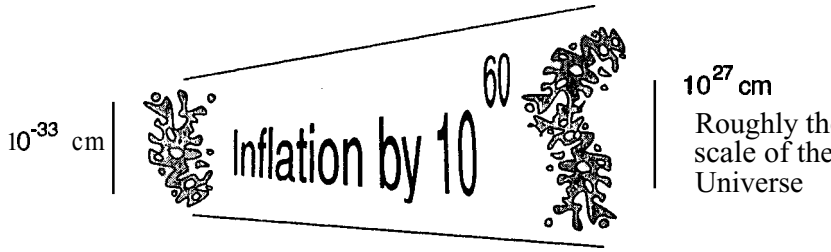



FIGURE 27. Illustrating the problem of the inflation of irregularities in the early Universe.

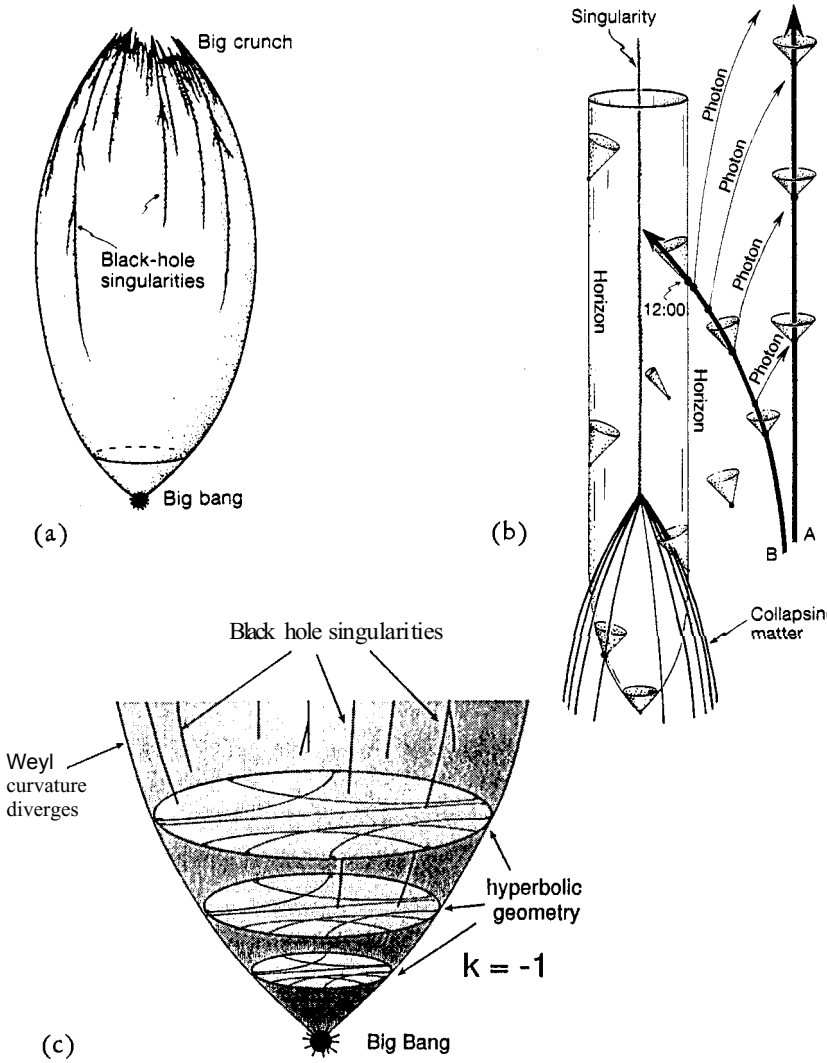



FIGURE 28. (a) The entire history of a closed universe that starts from a uniform low-entropy Big Bang with $Weyl=0$ and ends with a high-entropy Big Crunch —representing the congealing of many black holes — with $Weyl \rightarrow \infty$. (b) A space-time diagram depicting collapse to a black hole. (c) The history of an open universe, again starting from a uniform low-entropy Big Bang with $Weyl=0$.

perts. Matter collapsing into a black hole produces a singularity and that is what the dark lines on the space-time diagrams of the Universe represent. I want to introduce a hypothesis that I call the *Weyl curvature hypothesis.* This is not an implication of any known theory. As I have said, we do not know what the theory is, because we do not know how to combine the physics of the very large and the very small. When we do discover that theory, it should have as one of its consequences this feature that I have called the Weyl curvature hypothesis. Remember that the Weyl curvature is that bit of the Riemann tensor that causes distortions and tidal effects. For some reason we do not yet understand, in the neighbourhood of the Big Bang, the appropriate combination of theories must result in the Weyl tensor being essentially zero, or rather be constrained to be very small indeed.

That would give us a Universe like that shown in figure 28(c) and not like that in figure 28(a). The Weyl curvature hypothesis is time-asymmetrical and it applies only to the past type singularities and not to the future singularities. If the same rule of vanishing Weyl tensor that I have applied in the past also applied to the future of the Universe, in the closed model, you would end up with a dreadful looking Universe with as much mess in the past as in the future (figure 29). This looks nothing like the Universe we live in.

What is the probability that, purely by *chance,* the Universe had an initial singularity looking even remotely as it does? The probability is less than one part in $10_{10123}$ Where does this estimate come from? It is derived from a formula by Jacob Beckenstein and Stephen Hawking concerning black hole entropy and, if you apply it in this particular context, you obtain this enormous answer. It depends how big the Universe is and, if you adopt my own favorite Universe, the number is, in fact, infinite.

What does this say about the precision that must be involved in setting up the Big Bang? It is really very, very extraordinary. I have illustrated the probability in this cartoon of the Creator,

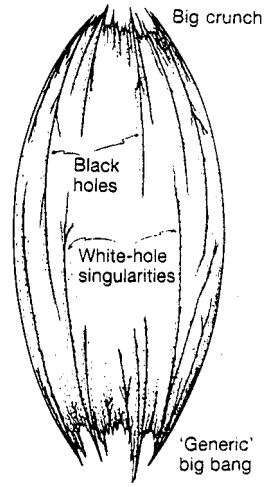



FIGURE 29. If the constraint $Weyl=0$ is removed, then we have a high-entropy Big Bang also, with $Weyl \to \infty$ there. Such a universe would be riddled with white holes, and there would be no Second Law of Thermodynamics, in gross contradiction with experience.

finding a very tiny point in that phase space that represents the initial conditions from which our Universe must have evolved if it is to resemble remotely the one we live in (figure 30). To find it, the Creator has to locate that point in phase space to an accuracy of one part in $10^{10^{123}}$. If I were to put one zero on each elementary particle in the Universe, I still could not write the number down in full. It is a stupendous number.

I have been talking about precision —how mathematics and physics fit together with extraordinary accuracy. I have also talked about the Second Law of Thermodynamics, which is often thought of as a rather floppy law —it concerns randomness and chance, and yet there is something very precise hiding underneath this law. As applied to the Universe, it has to do with the precision with which the initial state was set up. This precision must be something to do with the union of quantum theory and general relativity, a theory we do not have.

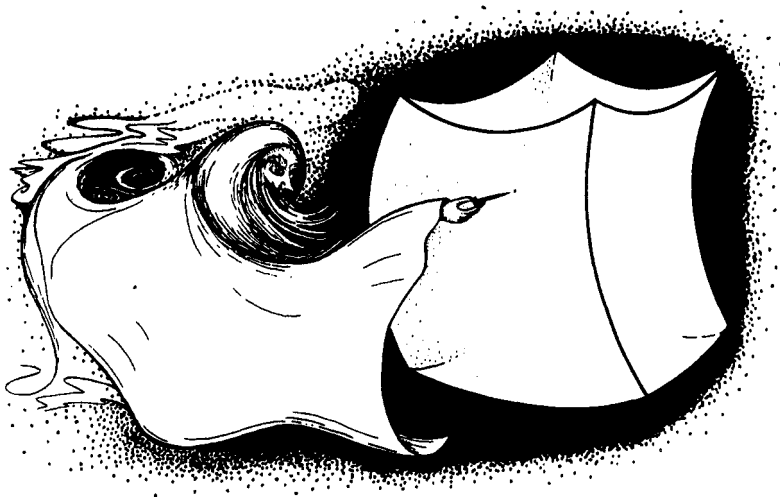


FIGURE 30. In order to produce a universe resembling the one in which we live, the Creator would have to aim for an absurdly tiny volume of the phase space of possible universes —about $1/10^{10^{123}}$ of the entire volume, for the situation under consideration. (The pin and the spot aimed for are not drawn to scale!)