Antoine Suarez
Peter Adams *Editors*

Is Science Compatible with Free Will?

Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience

Springer

Is Science Compatible with Free Will?

Antoine Suarez • Peter Adams
Editors

Is Science Compatible with Free Will?

Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience

Springer

Editors
Antoine Suarez
Center for Quantum Philosophy
Institute for Interdisciplinary Studies
Zürich, Switzerland

Peter Adams
Thomas More Institute
London, United Kingdom

ISBN 978-1-4614-5211-9 ISBN 978-1-4614-5212-6 (eBook)
DOI 10.1007/978-1-4614-5212-6
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012952284

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This book is the result of work initiated in a meeting organized by the Social Trends Institute (STI) and held in Barcelona at the IESE Campus-Nord from October 28 to 30, 2010.

As I said in my invitation to the participants, this conference aimed to discuss the idea that science today is compatible with phenomena governed by nonmaterial principles like, for instance, free will and consciousness. I would like to explain briefly how I came to organize this meeting and edit this book.

To begin with, I have the deep conviction that the three passions governing my life are compatible with each other: the desire for freedom, my religious faith, and science. I am not sure whether "a strong Faithful in the Church of the Larger Hilbert Space" fills that need to harmonize these three elements (see Chap. 4 of this book). As for me, it would be difficult to live were I to realize that in science there is no place for freedom or faith.

I believe that my existence cannot be explained exclusively by material principles: somehow I share in a nonmaterial, spiritual dimension. If I accept this, I have consequently to accept that the movement of my lips, my tongue, my eyes, when I am speaking to you, cannot be explained exclusively by a chain of temporal causes going back to the Big Bang. This means: one cannot claim to be a free being, or a believer, without intruding on scientific territory. Anyone who believes in God or a spiritual human soul cannot honestly claim that faith and science are two Non Overlapping Magisteria. On this point I agree with Richard Dawkins: Even rejecting any fundamentalism or creationism, as I do, one cannot help acknowledging that the domain of religion and that of science overlap to some extent. And if for you, both faith and science are vital, then you will conclude that a science excluding freedom and religion is likely not to be the last word in scientific knowledge.

The second part of my motivation for organizing the meeting that is the origin of this book has been decisively shaped by my encounter with nonlocality. After reading John Bell's "Essays on quantum philosophy" (1987) I had the intuition that the principle of nonlocality made possible what I was longing for: to be able to describe a world that can be governed by nonmaterial principles. John was not only

a "quantum engineer" but also a "quantum philosopher" (His wife Mary Bell used to joke that John would have very much liked to rent a flat in the so-called "Boulevard des philosophes" in Geneva). In CERN and in other research institutions I had the privilege of organizing with him some seminars on quantum philosophy from 1988 till 1990, the year he died. On the Internet you can find a video of one such seminar at CERN where John explains his famous theorem.

The discussions with John Bell inspired me to make the proposal for the *before–before* experiment, which I published together with Valerio Scarani. That was in 1997. Research on nonlocality was not main stream at that time. I was lucky to come into contact with a private Swiss banker in Geneva (Marcel Odier) who was ready to finance the experiment, and even luckier to encounter Nicolas Gisin and Hugo Zbinden who took on the challenge of performing the experiment. I will not enter into details here. Basically, we proposed a temporal explanation of nonlocality, much in the line of Bohm's theory (see Chaps. 3 and 5 of this book). And we expected to prove quantum mechanics wrong. As Hugo Zbinden says, it was probably the only experiment in his life where he thought quantum mechanics could be ruled out. I myself was even more convinced. Indeed Nicolas one day told me quite seriously that the data would be submitted to very strict checking before publication to avoid any mistakes coming from "wishful thinking." Nevertheless Nicolas himself used to say: "if the forthcoming results falsify quantum mechanics we will have enough work to last until the end of our days." It was a funny situation: Even wishing for a science where there is a place for spiritual (nonmaterial) action, I kept instinctively to a time-ordered causal explanation. On the morning of Friday 22nd June 2001 we attended the regular colloquium of Nicolas Gisin's group. André Stefanov presented the results, which confirmed quantum mechanics and refuted my temporal explanation: I thought I was assisting at my funeral. The story took an unexpectedly dramatic turn after lunch: André and I checked the apparatus and behold: one of the beam-splitters was wrongly oriented! The measurements had to be repeated during the next week. On Tuesday June 26th I realized that temporal causality is a preconception, a leftover of classic physics. I still remember the time: It was 19.15 h. And the verdict of the measurements some days later was clear: quantum mechanics prevails.

More recently, in 2010, I proposed a new experiment aiming to demonstrate other important implications of quantum mechanics related to the assumption that the decision of the experiment's outcome happens at detection (the so-called "collapse of the wave-function"). The experiment has been completed and published in May 2012, again in collaboration with Nicolas Gisin and his group (see Chap. 5, Sect. 5.3). The results demonstrate that the most fundamental principle ruling the material world, the conservation of energy, requires nonlocal coordination of detection outcomes, i.e., nonmaterial agency from outside spacetime. Additionally, the experiment is a natural and most direct demonstration of nonlocality in a context where the violation of Bell inequalities cannot be used as a criterion for establishing nonlocality (see Chaps. 3 and 5). In this sense, the experiment highlights the fact that the principle of nonlocality rules the whole of quantum physics and the material world emerges from nonmaterial features.

I wonder now why in 1997 I proposed the before–before (and expended considerable work, time, and money to do it) instead of proposing and doing the conceptually far more important and technically much less challenging experiment that I proposed in 2010 and has been done in the past months. A possible explanation may be that the new experiment is important not only because it is about nonlocality, but primarily because it demonstrates that nonlocality is crucial for the conservation of energy. To reach this insight, which now seems trivial to me, it was probably necessary to be defeated by quantum mechanics (in the field of the before–before experiment) after having very much expected to beat it. Now I really understand how important the quantum mechanical assumption of decision at detection is.

Through these and many other experiments in the past 10 years (see Chap. 3) we have reached a better understanding of what nolocality means: "that quantum correlations happen without the flow of time," "that quantum correlations come from outside spacetime," "that spacetime does not contain the whole of physical reality," "that quantum phenomena cannot be explained exclusively by material principles."

These insights were decisive for the project behind this book. I think it is not necessary to have the psi ability of "clairvoyance" to see that results proving that "quantum phenomena come from outside spacetime" and "conservation of energy requires nonmaterial agency" define a new era in science. In fact, they support the view that nonmaterial principles can steer the material world. So, during these years I was dreaming of bringing together neuroscientists, quantum-physicists, economists, and philosophers to reflect about this. This dream becomes fulfilled with the publication of this book. However, I would like to stress that my original insights and motivation are not necessarily shared by the other contributors, and each of them accepts responsibility only for the conclusions he or she draws. And for sure, this book represents only a first step towards promoting the understanding that "the world we see is made from things that are invisible," and I hope very much that this effort can be continued in the coming years.

I am enormously grateful to STI President Carlos Cavallé, Secretary General Tracey O'Donnell, Project Manager Fiona McCarthy, and all the other members of the Board and Staff of STI for sponsoring and organizing the Barcelona Experts Meeting, as well as all the participants at the meeting and contributors to the volume.

I am especially indebted to Peter Adams for his collaboration during the whole genesis of the project and in particular for assisting with the editing work. I acknowledge the enjoyable collaboration with Welmoed Spahr and Morgan Ryan from Springer during the production of the book.

Zürich, Switzerland, December 2012 Antoine Suarez

Foreword: The Social Trends Institute (STI)

The Social Trends Institute (STI) is an international research center dedicated to the analysis of globally significant social trends. By promoting research and scholarship of the highest academic standard within four subject areas—Family; Bioethics; Culture and Lifestyles; and Corporate Governance—STI aims to make a scholarly contribution towards understanding the varying and complex trends that characterize the modern world.

STI organizes Experts Meetings around specific topics within one of the research branches. These meetings are intended to foster open, intellectual dialogue between scholars from all over the world and from different academic backgrounds and disciplines. The scholars meet to present and discuss original research papers in an academic forum. These papers are then reviewed and edited in light of the conference discussion before publication.

This volume, *Is Science Compatible with Free Will? Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience*, is the result of one such Experts Meeting held in Barcelona in October, 2010 under the academic leadership of Antoine Suarez to explore the question "Is Science Compatible with Our Desire for Freedom?"

This query is particularly suited to STI's multidisciplinary approach. To fully explore and suggest solutions to the apparent conflict between deterministic science and the concept of human free will, STI gathered neuroscientists, physicists, and philosophers, as well as a biologist and an economist.

The results of their investigations are presented in this book. Without endorsing any particular viewpoint, STI hopes that as a whole, these contributions will deepen readers' understanding of this important question.

Tracey S. O'Donnell
Secretary General, Social Trends Institute
Barcelona, 2012

Contents

Contributors

Antonio Acín ICFO-Institut de Ciencies Fotoniques, Castelldefels, Spain

ICREA-Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain

Peter Adams Thomas More Institute, London, UK

Javier Bernácer Mind-Brain Project, ICS (Institute for Culture and Society), University of Navarre, Navarre, Spain

Gilles Brassard Département IRO, Université de Montréal, Montréal, QC, Canada

Luís Cabral Stern School of Business, New York University and Research Fellow, IME and SPSP (IESE) and Research Fellow, CEPR,

Robert O. Doyle Astronomy Department, Harvard University, Cambridge, MA, USA

Leonardo Fogassi Department of Neuroscience, University of Parma, Parma, Italy

Italian Institute of Technology, Rete Multidisciplinare Tecnologica, University of Parma, Parma, Italy

José Manuel Giménez-Amaya Mind-Brain Project, Research Group of Science, Reason and Faith (CRYF), ICS (Institute for Culture and Society), University of Navarre, Navarre, Spain

Nicolas Gisin Group of Applied Physics, University of Geneva, Geneva, Switzerland

Sara L. Gonzalez Department of Clinical Neuroscience, Electrical Neuroimaging Group, University of Geneva, NEUCLI, Geneva, Switzerland

Rolando Grave de Peralta Department of Clinical Neuroscience, Electrical Neuroimaging Group, Sleep Research Laboratory and Geneva Neuroscience Center, Geneva University Hospital, NEUCLI, Geneva, Switzerland

Martin Heisenberg Rudolf-Virchow-Centre, University of Würzburg, Würzburg, Germany

Jana M. Iverson Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA

Robert Kane The University of Texas at Austin, Austin, TX, USA

Flavio Keller Laboratory of Developmental Neuroscience, Università Campus Bio-Medico, Rome, Italy

Alfred R. Mele Florida State University, Tallahassee, FL, USA

Zeeya Merali Foundational Questions Institute, Decatur, GA, USA

Stephen Perrig Sleep Research Laboratory and Geneva Neuroscience Center, Service of Neurospychiatry, Geneva University Hospital, Geneva, Switzerland

Paul Raymond-Robichaud Département IRO, Université de Montréal, Montréal, QC, Canada

Giacomo Rizzolatti Department of Neuroscience, University of Parma, Parma, Italy

Brain Center for Social and Motor Cognition, Italian Institute of Technology, Rete Multidisciplinare Tecnologica, University of Parma, Parma, Italy

Jean Staune Université Interdisciplinaire de Paris, Paris, France

Antoine Suarez Center for Quantum Philosophy, The Institute for Interdisciplinary Studies, Zurich, Switzerland

Social Trends Institute/Bioethics, Barcelona, Spain

Giulio Tononi Department of Psychiatry, University of Wisconsin, Madison, WI, USA

Chapter 1
Introduction

Peter Adams and Antoine Suarez

Abstract This is the first book to discuss, at the same time, the implications of quantum physics, Libet's experiments and the neurophysiological finding of mirror neurons, for consciousness, interpersonal communication and our desire for freedom. The authors present perspectives coming from different disciplines, ranging from those focusing on the scientific background, to those highlighting rather more a philosophical analysis. However, all the contributions share a common characteristic: They take current scientific observations and data as the basis from which to draw philosophical implications.

Keywords Determinism • Libet's experiments • Quantum non-locality • Mirror neurons • Non-material principles • Free will • Limited consciousness

Anyone who claims the right "to choose how to live their life" excludes any purely deterministic description of their brain in terms of genes, chemicals or environmental influences. When you claim to be the author of a paper and to express your own thoughts, you assume that, in typing the text, you govern the firing of the neurons in your brain and the movement of your fingers through the exercise of your own free will: What you write is not completely pre-determined at the beginning of the universe.

P. Adams (✉)
Thomas More Institute, 18b Netherhall Gardens, London NW3 5TH, UK
e-mail: peter.adams@thomasmoreinstitute.org.uk

A. Suarez
Center for Quantum Philosophy, The Institute for Interdisciplinary Studies,
Berninastr. 85, 8057 Zurich, Switzerland

Social Trends Institute/Bioethics, Barcelona, Spain
e-mail: suarez@leman.ch

A. Suarez and P. Adams (eds.), *Is Science Compatible with Free Will?: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience*,
DOI 10.1007/978-1-4614-5212-6_1,

"We experience ourselves as free mental beings but the scientific view does not admit any room for a mental agent like free will, which influences neurons and produces actions [...]. When I observe the brain I cannot find any evidence of a mental agent like free will or personal responsibility. Nevertheless when I get home in the evening I hold my children responsible for their actions if they have done any nonsense". These words of the German neuroscientist Wolf Singer describe well the deep conflict between the conviction of daily life that a human being somehow shares in freedom and responsibility, and the description of the human brain provided by the prevailing deterministic neuroscience.

When faced with this conflict two alternative positions are possible: Either human freedom is an illusion, or deterministic neuroscience is not the last word on the brain, and will eventually be superseded by a neuroscience that admits processes not completely determined by the past.

This book aims to investigate whether it is possible to have a science in which there is room for human freedom.

The arrival of quantum physics replaced the deterministic view of the world. In this sense quantum randomness is in principle good news for "free will." However, a frequent objection to the possible relevance of quantum physics for the question of free will is that quantum non-deterministic randomness is just as bad as determinism: "If nature is fundamentally random, then the outcomes of our actions are also completely beyond our control". In the end, it seems that neither determinism nor randomness is good for free will.

Another big objection to free will comes from present neuroscience and mainly from the experiments of Benjamin Libet. On the basis of these experiments one often states that our feeling that we make conscious decisions is an illusion. At the instant we are aware of a decision, the neural activity responsible for such a decision arose unconsciously in the brain prior to that awareness.

These two main objections are discussed by contributors to this volume.

Firstly, recent experiments suggest that there is no incompatibility between quantum randomness and freedom. What is more, today's quantum physics highlights that there are observable effects which cannot be explained by any narrative in space-time and, in this sense, come from outside space-time. This could offer a framework for a description of the world that does not exclude immaterial agency in principle and therefore remains open to ideas and concepts like freedom, personal identity, creativity, responsibility and religious faith. Contributions in this book stress that, for the sake of freedom, quantum non-material agency coming from outside space-time is a much more relevant concept than quantum randomness.

Secondly, the book's contributors offer three responses to Libet's experiments. There are objections to the experimental protocols themselves. It is also argued that Libet accepts the possibility of a conscious veto after the arousal of the neural readiness potential, capable of stopping the action. Therefore, the final action does not abolish responsibility. In addition, the idea of voluntary unconscious movements may contribute to making Libet's results compatible with free will and moral responsibility. Indeed one can interpret Libet's results as a demonstration

that voluntary (non-deterministic) movements are not necessarily conscious ones. At the moment the subject agrees to start the experiment, he makes a free and conscious decision. By contrast when the experiment is running and he has to decide to move his wrist, this decision may occur in an unconscious way. This last interpretation is very much related to the medical praxis of considering the presence of spontaneous voluntary movements, like breathing or eye movements in a PVS patient or a child with hydranencephaly, as a clear sign that the person is not dead. Libet's experiments actually support the idea that human consciousness is limited and voluntary actions in humans can exhibit many degrees of consciousness, going from unconscious voluntary actions to highly conscious ones.

Supporters of free will among the contributors to the book state that free will is an axiom: it cannot be proved or rejected by any scientific experiment. You can choose to do deterministic science, as for instance in the many-worlds interpretation. However, as a matter of fact, standard quantum physics admits, as an axiom, that the experimenter is free, and quantum physics is quite successful as a science!

Additionally, science is discovering quantum interference behind life phenomena. Science-based speculation allows us to establish a correspondence between quantum interference and the way the brain functions. The output of the brain, like that of a quantum interferometer, may exhibit coordination coming from non-material agency. In an unconscious state (as for instance during certain periods of sleep or in a PVS patient) this coordination is at a very low level, similar to that happening in a quantum device in the lab, and the outcomes fulfil a certain statistical distribution depending on the physiological parameters in the brain. What characterizes the state of consciousness is the capability of self-influencing these parameters. Sense data is also able to set these parameters. So that when I perceive some behaviour outside of me, the neurons which become activated in my brain may be the same as those which become activated when I perform the same actions myself (mirror neurons).

The work leading to this book was initiated by a group of researchers (neuroscientists, quantum physicists and free-will philosophers) convened by the Social Trends Institute (STI) to study this topic. The meeting was held in Barcelona at the IESE Campus-Nord from October 28–30, 2010. Most of the chapters in this volume are papers presented at the STI Meeting and have been subsequently updated and revised. Some chapters were commissioned and added later on.

In the Chapters 2–5 Antonio Acín, Nicolas Gisin, Gilles Brassard and Paul Raymond-Robichaud and Antoine Suarez present recent experiments in quantum physics on randomness and non-locality, and discuss the implications of these results for free will and non-material agency. While Acín, Gisin and Suarez support non-locality and free will as fundamental principles of quantum physics, Brassard and Raymond-Robichaud present a new alternative local and deterministic view called "parallel lives".

In the Chapters 6–7 Zeeya Merali and Martin Heisenberg discuss the possibility that free will is not only relevant for human decisions but also for explaining non-human nature, and in particular the behaviour of animals.

The Chapters 8–11 present basic results of neuroscience. Flavio Keller and Jana M. Iverson discuss the basic role that voluntary inhibition in humans may have as a

prerequisite for the emergence of free will. Leonardo Fogassi and Giacomo Rizzolatti explain their famous discovery of "mirror neurons" and its relevance for understanding action and intention, the very basis of interpersonal relationships and communication. Sara L. Gonzalez, Stephen Perrig and Rolando Grave de Peralta address the crucial question of distinguishing conscious and vegetative state. Giulio Tononi presents his integrated information theory of consciousness, according to which a conscious choice, "while maximally and irreducibly causal, is also necessarily under-determined and thus unpredictable".

The Chapters 12–15 argue that neuroscience, and in particular Libet's experiments, do not provide sufficient reason to deny free will. Javier Bernácer and José Manuel Giménez Amaya study the neurophysiological basis of "habits" and comment on Libet's experiments. Alfred Mele proposes a new interpretation of Libet's results, which is compatible with free will. Jean Staune stresses that Libet himself admitted that the subject always remains free to "veto". Luis Cabral discusses "free will" with relation to the emerging approach of neuroeconomics. The idea that, if we are able to measure brain activity well enough, then economic behaviour will be predictable and can be used as a "platform" for a theory of deterministic human behaviour. Cabral disagrees with this view and believes there is an irreducible degree of uncertainty which results from each individual's free will.

In the Chapters 16 and 17 Bob Doyle and Robert Kane make proposals about how to reconcile indeterminism and free will without reducing this to mere chance or mystery.

Finally, in Chapter 18, Peter Adams and Antoine Suarez try to show how the different arguments presented in this book, while coming from quite different and apparently disparate disciplines, are related and complement each other. In addition, they point out important philosophical challenges coming from science that we will have to tackle in the coming decades.

In summary, today's scientific view seems to admit room for mental agencies involving free will and consciousness, which influence neurons and produce actions. This view fits well with those intuitions and convictions of daily life that (as referred to above) even deterministic neuroscientists confess to having: "we experience ourselves as free mental beings" and "when I get home in the evening I hold my children responsible for their actions if they have done any nonsense". Thus it seems possible to reconcile modern science with our innermost desire for freedom.

This is the first book to discuss, at the same time, the implications of quantum physics, Libet's experiments and the neurophysiological findings of voluntary inhibition, mirror neurons, vegetative state and sleep, for consciousness, interpersonal communication and our desire for freedom. As already said, the authors present perspectives coming from different disciplines, and range from those focusing on the scientific background, to those highlighting rather more a philosophical analysis. However, all the contributions share a common characteristic: They take current scientific observations and data as the basis from which to draw philosophical implications. It is these features that make this volume unique, an exceptional interdisciplinary approach combining scientific strength and philosophical profundity. We are convinced that it will strongly stimulate the debate and contribute to new insights in the mind–brain relationship.

Part I
Quantum Physics and Free Will

Chapter 2
True Quantum Randomness

Antonio Acín

Abstract Randomness is a fascinating concept. Since the early days of quantum physics, it became clear that a new form of randomness, with no classical analogue, appears in the quantum regime. Still, it has only been recently that tools to certify and quantify the presence of intrinsic quantum randomness have been introduced. These tools have also been exploited to certify the generation of randomness in an experiment involving two distant atoms. In this contribution, we review the novel approach to quantum randomness and discuss the main differences and advantages when compared to the existing approaches, both in the classical and quantum regime.

Keywords Randomness • Quantum physics • Bell's theorem • Quantum information theory

2.1 Introduction

The concept of randomness has attracted and keeps attracting the interest of many different scientific communities. From a fundamental point of view, a crucial question in physics (and even philosophy) is whether nature is deterministic or intrinsically random. Strictly speaking, there is no such a thing as true randomness in the classical world: randomness is simply a consequence of lack of knowledge. Indeed, in any physical situation involving different particles, if an observer has a complete description of the initial positions and velocities of the particles and the interactions among them, he can predict with certainty the status of these particles at any given time. Although this may be an extremely difficult task, as it may

A. Acín (✉)
ICFO-Institut de Ciencies Fotoniques, 08860 Castelldefels, Spain

ICREA-Institució Catalana de Recerca i Estudis Avançats, 08010 Barcelona, Spain
e-mail: antonio.acin@icfo.es

A. Suarez and P. Adams (eds.), *Is Science Compatible with Free Will?: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience*,
DOI 10.1007/978-1-4614-5212-6_2,

require unlimited computational capabilities, perfect predictability is in principle possible. This (classical) determinism can be found, for instance, in the introduction to the *A Philosophical Essay on Probabilities* by Pierre-Simon Laplace (Laplace 1840):

> We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes.

True randomness can however be found in the quantum world: even if one has perfect knowledge about the preparation of the state of a quantum system, there are experimental observations for which quantum theory can only predict the answer in terms of probabilities. That is, now randomness is not a consequence of lack of knowledge but, in a way, a limitation on the predictability of the theory. Probably, this is one of the reasons why Albert Einstein disliked the quantum formalism ("God does not play dice"). Indeed, he believed in the existence of another theory alternative to quantum physics in which this new form of randomness could again be understood as a consequence of ignorance, as in the classical setting. This alternative theory, proposed in Einstein et al. (1935) together with Podolsky and Rosen, would be more complete than quantum physics, in the sense that it would contain new variables which are not present in the quantum formalism (often these variables are named "hidden"). Knowing these, at the moment hidden, variables would allow recovering the determinism of our physical description of nature. In 1964, however, John Bell proved that this alternative theory based on hidden variables would lead to experimental predictions which are in conflict with quantum physics (Bell 1965, Bell 2004). More precisely, he showed that the correlations observed between the results of measurements applied on some quantum states of two particles cannot be reproduced by these models.[1] This result, known as Bell's Theorem, paved the way to the experimental falsification of these alternative hidden-variable theories, which was for instance accomplished in 1982 by Aspect and co-workers (Aspect et al. 1982). From the point of view of randomness, the experimental falsification of these alternative theories confirms the existence of a new form of randomness in the quantum world.

Beyond all these fundamental issues, randomness is also an extremely valuable resource in our society with applications in many different areas (Knuth 1981). Random numbers are constantly used for cryptographic applications, gambling or numerical methods for the simulation of physical and biological systems. Due to their relevance, there is an intensive ongoing effort to (i) develop good sources of

[1] Actually, determinism can still be recovered if the hidden variables allow faster-than-light communication. But this would in turn be in conflict with Einstein's Special Relativity! We come back to this point below.

random numbers and (ii) design reliable tests to certify the random nature of the generated numbers. These two issues are strongly connected as the quality of a source is estimated using the tools developed for certification. Randomness certification is indeed a crucial question and, as discussed below, it is notoriously difficult to ascertain with the existing techniques the random properties of a device.

2.2 Motivations and Main Results

Nowadays, there basically exist three types of random number generators (RNG): "true" RNG, pseudo-random number generators and quantum RNG. In the first case, some initial numbers are generated by means of a physical process that is hard to predict, such as the noise in electrical circuits or the timing of user processes. Pseudo-random number generators consist of the output of a deterministic function applied to a shorter seed, assumed to be random and possibly produced by a true RNG. Finally, quantum RNG use quantum features for the generation of the random numbers. However all these solutions suffer from the following three drawbacks, which are relevant both from a conceptual and applied point of view.

The first problem concerns the issue of *randomness verification*. Although all these different approaches to randomness generation are based on different principles, they all use the same framework to certify the randomness of the produced numbers: it is always measured by a series of statistical tests (Marsaglia 2008, Rukhin et al. 2008) designed to check the absence of patterns in the generated sequences. It is however unclear what passing all these tests mean from the point of view of randomness. No finite set of tests can be considered complete (Rukhin et al. 2008), since it can never be excluded the existence of patterns that are not covered by the existing battery of tests. In particular, the tests should be periodically revaluated and, if needed, corrected.[2] Thus, it is highly non-trivial, if not impossible, to ascertain with the existing techniques the random character of an experiment.

Second, many applications, especially for cryptographic purposes, require *private randomness*. In these applications, the randomness, or unpredictability, of the generated events is exploited to achieve a given task, such as secure information transmission. This requires that the numbers generated by the device should not only appear random, in the sense of hard to predict, to the honest user, but also to any other, potentially adversarial, user. Unfortunately, the reliability of the statistical tests is even less clear in these applications. For instance, systematic errors in the generators can introduce patterns that may be undetected by the statistical tests applied by the honest user, but predicted by a computationally more powerful adversary. These patterns can then be exploited to break the cryptographic protocol.

[2] One of the most famous examples of bad RNG was RANDU, a RNG used in the 60–70s which was later discovered to have a well-defined pattern, see http://en.wikipedia.org/wiki/RANDU#cite_note-Entacher-2000-0.

Third, the situation becomes more critical in the non-trusted provider scenario, where the devices used for the generation cannot be trusted and should be seen as a black box generating the numbers. In this scenario, there cannot be any classical technique proving the presence of private randomness. Indeed, one can never exclude, for instance, that the numbers have been prepared in advance by an adversary using a very "good" RNG. These numbers have been copied into a memory inside the device and then, despite looking random, can be completely predicted by this adversary. In order to avoid this problem, the proposal for randomness generation should be *device-independent*: the random properties of the generated numbers should not rely on any modelling of the internal working of the devices used in the generation. The device-independent property provides a second advantage for randomness generation: as the scheme does not depend on the internal working of the devices, it is more robust to preparation imperfections or drifts during the generation process.

Finally, there is a fourth issue which only concerns the existing quantum solutions. It seems quite unsatisfactory, and even contradictory, to verify their quality by means of the same techniques used for classical devices, which are always of deterministic nature. It is hard to claim that an intrinsic quantum property has been used for randomness generation if this is certified by tests that are also satisfied by classical devices. It would then be desirable to derive new forms of randomness certification based on quantum principles with no classical analogue. This is intimately related to the fact that, although quantum physics contains an intrinsic form of randomness, in any real situation this randomness is necessarily mixed up with an "apparent" randomness that results from noise in the system or lack of control of the experiment. In order to disentangle these two forms of randomness, one should derive *tools to detect and quantify the intrinsic quantum randomness* generated in an experimental setup.

In Pironio et al. (2010), we present a completely novel approach to randomness generation in which all these problems can be solved. First, we establish a fundamental link between the correlations between quantum particles and randomness. This link allows for the first time to quantify the intrinsic quantum randomness in an experimental setup, which can now be disentangled from any apparent randomness associated to imperfections or lack of knowledge. Then, we show how our techniques can be used to design a new type of RNG leading to numbers which are (a) certifiably random, (b) cryptographically secure and (c) device-independent. Finally, we illustrate how our techniques can be applied in a real setup and prove that 42 new random bits have been generated in an experiment involving two distant ions.

2.3 Context and Existing Results

Before moving to the more detailed discussion of our results, we describe the background and tools used for their derivation, and mention the existing results. First of all, our results represent a novel application of quantum information theory

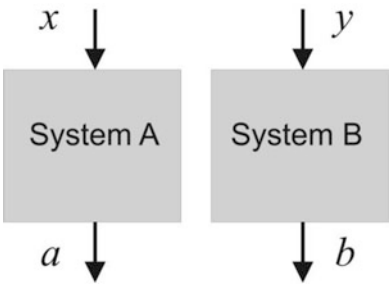


Fig. 2.1 *Schematic representation of the device-independent scenario*. In the device-independent scenario, several users have access to uncharacterized boxes (two in the figure), possibly prepared by an adversary. The users can choose an input for the boxes and get an output as a result. This scenario is exactly the same as in any Bell test. In the case of randomness generation, there is a single user who has access to two separated boxes. The user should check that the boxes produce a Bell inequality violation and construct the list of random bits from the generated outputs

(Nielsen and Chuang 2000). This is a highly inter-disciplinary field which combines concepts and techniques from physics, mathematics, computer science and engineering. The main goal is to understand how the quantum formalism can be exploited for information processing and communication. Remarkably, novel applications, such as secure cryptography (Bennett and Brassard 1984) or potentially more powerful computers (Shor 1997), become possible thanks to quantum effects.

In the last years, some Quantum Information applications have been moved to the device-independent scenario, where the goal is to design protocols achieving an information task without making any assumptions on the internal working of the devices used in the protocol. This makes the protocols (i) more robust against imperfections and drifts on the devices and (ii) opens the field to the non-trusted provider scenario, in which the devices may have been prepared by an adversary. The device-independent scenario consists of different users who have access to devices, which are seen as black boxes producing a classical output given a classical input (see Fig. 2.1). After testing the devices, the users can characterize their statistical properties, that is, they can infer the probabilities of obtaining all the possible outputs for any combination of the inputs. From this statistical description, the users should conclude whether an information task can be achieved. However, no assumption is made on the process generating the outputs of the devices given the inputs, apart from the fact that it should not contradict any quantum law. Initially proposed in the framework of quantum key distribution (Acin et al. 2007, Pironio et al. 2009), the device-independent scenario has been extended to other problems and, at the moment, represents one of the most active areas in quantum information theory (Ekert 2009). Our work naturally fits entirely into this picture: it exploits quantum laws to solve an information task which is impossible within Classical Information Theory, namely the generation of certified private randomness, in the device-independent scenario.

The key quantum property behind our proposal is quantum non-locality. As mentioned in the introduction, John Bell proved that the predictions of the

classical hidden-variable models proposed by Einstein, Podolsky and Rosen (EPR) in Einstein et al. (1935) were in conflict with Quantum Physics. In particular, he proved that EPR models imply some experimentally testable conditions, known as Bell inequalities, which are violated by the observed correlations between the results of measurements performed on two separate quantum particles. This form of quantum correlations with no classical analogue is named quantum non-locality. From a fundamental point of view, quantum non-locality represents the most intrinsic and striking quantum property. Remarkably, it turns out to be also a useful resource for information applications: it plays a fundamental role for quantum communication complexity applications (Buhrman et al. 2010) and is an essential pre-requirement for device-independent protocols.

In a way or another, the relation between randomness and quantum non-locality has appeared in several previous publications. However there are two works which are clearly connected with our findings. First, our main intuition is related to Ekert's proposal for secure quantum key distribution using quantum non-locality (Ekert 1991). In Ekert's scheme, the security of the key is guaranteed by the fact that there cannot exist a hidden-variable model reproducing quantum measurement outcomes which violate a Bell inequality. The connection between randomness and quantum non-locality was strengthened by Colbeck in his PhD Thesis (Colbeck 2007). There, he proposed to use quantum non-locality for the task of private randomness generation. Inspired by all these ideas, we provide in Pironio et al. (2010) the first tools to connect quantum non-locality and randomness. We show (i) how to quantify the randomness contained in quantum correlations which lead to a Bell inequality violation, (ii) how this can be incorporated into a protocol for private randomness generation in a device-independent manner and (iii) how our techniques can be applied to a real experimental setup.

2.4 Statement of the Obtained Results

In this section we review the main results derived in Pironio et al. (2010). Here, the results are stated, while a more detailed explanation is provided in Appendices A, B and C. Some knowledge of quantum physics is needed to fully understand these appendices. Those readers who are not interested in the technical details can skip them.

Quantum non-locality and randomness: Our first result consists of a link between randomness and the violation of Bell inequalities. As mentioned previously, Bell inequalities are conditions satisfied by all models *à la* EPR. From a more operational point of view they also define limits on the way two separated devices can be correlated by means of classical instructions. These inequalities can be violated by the results of measurements performed on systems of two quantum particles. It is well established that this violation implies that a novel form of correlations without classical analogue becomes possible in the quantum framework. Interestingly, we also show how the violation of these inequalities

can be used to certify the presence of randomness, as it is possible to derive bounds on the amount of randomness produced in a quantum setup from the observed Bell violation.

The scenario is the same as in Fig. 2.1. Two separated observers perform different measurements, labelled by x and y, on two quantum particles and get measurement results a and b. By repeating this process, they can estimate the joint probability distribution, $P(a,b|x,y)$, of getting result a and b when they applied measurements x and y. From this distribution, the observers can compute the violation of a Bell inequality. If a violation is observed, then they can guarantee that the observed outcomes have some randomness. Actually, we can establish a quantitative link between the observed Bell violation and the amount of randomness. Our findings show that the more the particles are quantumly correlated (in the sense of violating a Bell inequality), the more random the measurement outcomes are. That is, they constitute a fundamental link between non-local quantum correlations and randomness, two of the main intrinsic and counter-intuitive properties of quantum physics.

Device-independent quantum random number generator: Next, we show how the previous bounds can be used to realize a new type of quantum random number generator (QRNG). As mentioned, and contrary to all previous solutions, our scheme produces randomness which is (a) certifiable, (b) private and (c) device-independent. It is based on a previous proposal by Colbeck (2007).

It is useful in what follows to work in the non-trusted provider scenario and, thus, assume that the user gets two devices from a non-trusted provider. Using these devices, the user should be able to perform a Bell test, as explained before and shown in Fig. 2.1. The final string of perfectly random bits will be made out of N pairs of results, $(a_1,b_1,\ldots,a_N,b_N)$, obtained by N uses of the devices. The random character of the generated numbers is guaranteed by the violation of a Bell inequality. Importantly, this holds true in a scenario where the internal workings of the device are not known, and even if the devices were prepared by an external agent who should not be able to bias or predict the random bits. This follows from the previous analysis: whatever the adversary prepared in the device for generating the output given the input, if it violates a Bell inequality, then there is a bound on the randomness of the outputs. For instance, a memory attack, in which the provider has generated in advance and copied the numbers into memories located in the devices, is impossible, as this would represent an EPR model for the measurement outcomes, which is impossible because of the observed Bell violation.

There is however an important point: the user of the devices does not know a priori whether the devices violate the Bell inequality and must estimate this using a statistical test. But the estimate cannot be carried out in a predetermined way. Indeed if the measurement settings that are going to be used are known in advance, then the external agent may have prepared a device that is completely deterministic, but which is such that on the specific sequence of inputs that are going to be used it appears to violate a Bell inequality. There is thus an apparent contradiction between the aim of making a random number generator and the requirement of using random settings to test the device. It is however possible to carry out the statistical test using

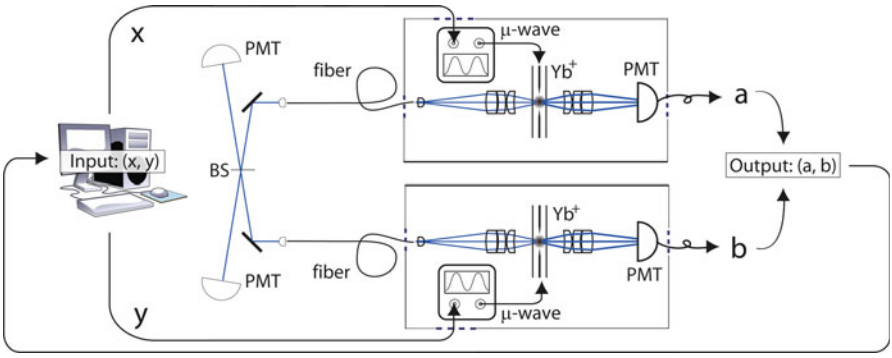


Fig. 2.2 *Experimental setup*. The figure shows the schematic representation of the experimental setup. The two particles in the two separated traps correspond to the two devices in Fig. 2.1. The choice of measurements is done by microwave pulses impinging the particles. The outcomes can take two possible values, corresponding to whether fluorescence light is detected at the detectors

only a very small amount of randomness, much smaller than the amount of randomness generated by the measurements. Thus non-trusted devices that violate a Bell inequality can be used as *randomness expanders* in which a small random seed is expanded into a much longer random string.

Finally, one could naively think that the construction of this device would automatically follow from the bounds derived in the present section. As explained in the appendix, this derivation is less straightforward than initially expected: in the non-trusted provider scenario, the devices may change in time and vary their response according to what was done before (e.g. keeping track of all the previous uses of the devices in a memory). All these effects cannot be described by the probability distribution $P(a,b|x,y)$, but can be taken into account using other theoretical tools. Thus, using these tools and the observed Bell violation, it is possible to derive a list of provably perfect random bits from the string of measurement outcomes $(a_1,b_1,\ldots,a_N,b_N)$.

Experimental generation of private random numbers: Finally, we performed a proof-of-principle experiment of our theoretical formalism together with the group of Prof. Monroe, at the University of Maryland. In this group, they can observe a Bell violation between two atoms located in two separated traps (Matsukevich et al. 2008), see Fig. 2.2. These traps should then be seen as the physical realizations of the abstract boxes in Fig. 2.1. The different measurement can be chosen by sending different microwave pulses into the atoms. The measurement results have two outcomes, which correspond to whether the atom emits fluorescence light back after the pulse.

In the experiment, data were recorded over a period of approximately one month to observe a violation of the simplest Bell inequality, namely the Clauser–Horne–Shimony–Holt (CHSH) inequality (Clauser et al. 1969). From the observed violation, and using the previous theoretical tools, we could certify that 42 random bits were produced in the quantum setup. Admittedly, the generation rate was not at all

competitive when compared to any of the existing random number generators. But our analysis certified that, for the first time, new intrinsic quantum randomness was produced in an experiment without a detailed model of the devices.

2.5 Randomness beyond Quantum Theory

There is no doubt that randomness is a fascinating concept which challenges our scientific understanding of nature. In our work, we provide a novel approach to the problem of randomness characterization and generation and show how quantum properties can be exploited to generate certifiably quantum private randomness in a device-independent manner. As mentioned, none of these crucial properties were met in any of the existing solutions, both classical or quantum. Clearly, our work has mostly an operational approach and exploits the link between Bell violation and randomness within quantum theory to solve an information problem, namely randomness generation. Still, this link has profound implications that go beyond the quantum formalism. In what follows, we abandon the operational approach and adopt a more speculative motivation to discuss some of these implications.

First, it is important to recall a fundamental relation, derived by Valentini (2002), among randomness, non-locality and the no-signalling principle. Before presenting it, let us recall that the no-signalling principle is probably the most accepted principle in physics. It states that information does not propagate instantaneously. That is, an action performed at a given location cannot have a noticeable effect in an arbitrarily distant location after a given amount of time, T. As the principle assumes that there exists a finite speed of information propagation, denoted by v, only those locations whose distance is smaller than vT can notice the effect of the previous action. Any existing physical theory, including Einstein's relativity or quantum physics (and any of its further developments), is compatible with this principle. Actually, even if Einstein's Relativity was proven to be wrong, in the sense that there are particles (or quasi-particles) travelling faster than light, the no-signalling principle only requires that this velocity is finite. In a way, it is just a consequence of the fact that energy cannot be unbounded, or, more in general, of the belief that there is a finite limitation for any physical effect.

What Valentini (and possibly others before and after him) showed is that:

Non-local correlations + Determinism → Signalling

In deriving this relation, no quantum property is needed; it is a general implication valid for any physical theory. Thus, if non-local correlations are experimentally observed in nature, one should abandon either determinism or the no-signalling principle (or both!). It is just a matter of choice. For instance, Bohm's Theory (Bohm 1952a, b) is a theory alternative to quantum physics with no randomness. However, if the theory was correct, it could be used to signal from one location to another instantaneously. Although it may seem surprising from outside the community, most physicists prefer to abandon determinism. This is simply because, as said, the no-signalling principle is possibly the best accepted and most natural

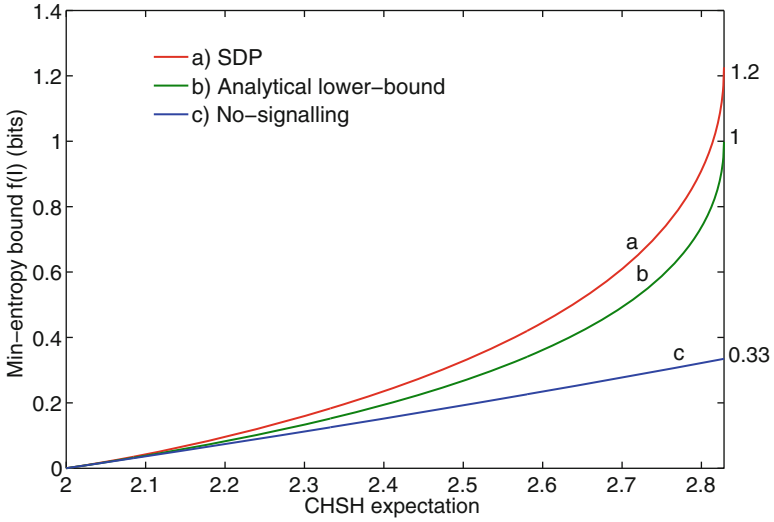


Fig. 2.3 *Bound on the randomness of the outcomes as a function of the CHSH inequality violation*. The plot shows the bound on the randomness of the outcomes for the case of the CHSH inequality. The point of no violation corresponds to a value of the CHSH expectation equal to 2. At this value, no randomness can be guaranteed, as expected. Curve a is computed numerically using semi-definite programming (SDP) techniques and gives the bound on the randomness of the two outcomes. Curve b corresponds to a single outcome and can be computed analytically. The point of maximal quantum violation is equal to $2\sqrt{2} \approx 2.8$. At this value, any of the two outcomes gives one perfect random bit, while some randomness is still left in the other outcome, as the red curve is slightly larger than 1.2 bits. The same bounds are computed in the no-signalling case (see discussion in the main text), corresponding to curve c. In this case, there is no difference between the randomness of one or the two outcomes, and it reaches a maximum value of 1/3

principle to any physicist. But, it is important to keep in mind that, in order to prove randomness from an observed Bell violation, the validity of the no-signalling principle is always implicitly assumed.

In the previous analysis, the validity of the entire quantum formalism was assumed. As this theory does not allow any form of signalling, the previous implication by Valentini automatically applies. In our case, we could go beyond this qualitative implication and derive quantitative tools to certify and quantify the presence of randomness in a real experimental setup. Still, similar techniques and bounds can be derived in a more general framework in which the validity of quantum theory is not assumed. Indeed, just the validity of the no-signalling principle suffices to derive bounds on the amount of randomness form an observed Bell violation (see also Appendix A and Fig. 2.3).

We conclude, then, from the previous discussion that the experimental observation of non-local correlations, together with the no-signalling principle, implies that Nature is random. But now, have non-local correlations really been observed? To be honest, a proper Bell inequality violation has never been observed. All the existing experimental Bell violations suffer from technological problems that do

not allow excluding a deterministic and no-signalling explanation for the observed data. In other words, exploiting the imperfections in the devices, one can construct ad hoc EPR models reproducing the measurement results. These models are highly artificial (for instance a photon deciding to produce a click in a detector depending on which measurement is applied) and have to be changed from experiment to experiment. But, in view of all the previous fundamental implications, it would be highly desirable to have a loophole-free Bell experiment. Moreover, it is also a relevant question from a practical point of view, as Bell inequalities can be exploited to solve important information tasks, such as randomness generation or secure key distribution.

Leaving aside the somehow artificial loopholes, which seem not to have a fundamental origin and simply be caused by technological limitations, what is really needed for the proper observation of non-local correlations? That is, assuming perfect technology, can the presence of non-locality be strictly proven? Here, as above, it is necessary to assume that the choice of the measurements at the devices, x and y, is random. In the most extreme case, this is often attributed to the free will of the observers, which can freely choose the measurements to be applied. Putting all these things together, and assuming that the loopholes will be closed and a proper Bell violation will be observed, we have

Free Will + Determinism $\rightarrow$ Signalling

That is, in a scenario in which observers are assumed to have free will and where instantaneous communication is impossible, the observation of non-local correlations implies the randomness of the outcomes.[3]

Finally, let us go one step further. Is it possible to guarantee the presence of randomness from other physical principles, without resorting to some initial seed of randomness, or without invoking free will? That is, can randomness be proven "from scratch"? Probably the answer to this question is negative. In any case, our quantitative study sheds light onto it. Indeed, one could naively argue that all the randomness seen in the measurement results is a consequence of the initial assumed randomness or free will. However, our expansion results prove that this is impossible: new non-previously existing randomness is generated by the quantum setup.

Appendix A: Quantum Non-Locality and Randomness

In this appendix we explain how to derive the link between randomness and the violation of Bell inequalities. The scenario is the same as in Fig. 2.1: two separate devices generate classical outputs a and b given the inputs x and y. By testing

[3] This is sometimes named as the Free-Will Theorem, see J. Conway and S. Kochen, *The Free-Will Theorem*, Foundations of Physics 36:1441–1473 (2006). But, as shown, this Theorem can simply be seen as a corollary of Valentini's implication.

the devices, it is possible to infer the probability distribution $P(a,b|x,y)$ describing the input–output relation. Since the devices are assumed to be quantum, this distribution has to be such that there exists a quantum state ρ and measurement operators for each device, M_a^x and M_b^y, reproducing it through the standard Born rule,

$$P_Q(a,b|x,y) = tr\left(\rho\, M_a^x \otimes M_b^y\right). \tag{2.1}$$

The tensor product in this equation follows from the fact that no interaction between the devices is assumed when the measurements take place. All the distributions compatible with condition (2.1) define the set of quantum correlations.

In order to derive a Bell inequality, one considers linear combinations of the input–output probability distributions, specified by a vector of real coefficients c_{ab}^{xy},

$$\beta = \sum_{a,b,x,y} c_{ab}^{xy}\, P(a,b|x,y). \tag{2.2}$$

For some of these coefficients, this expression (i) is bounded by β_L for EPR models, which defines the Bell inequality $\beta \leq \beta_L$, while (ii) there exist quantum states and measurements leading to a larger value. It is then said that these states and measurements violate the Bell inequality.

The standard measure of randomness in information theory is the min-entropy: for a probability distribution $P(z)$ describing a random variable Z which can take d possible values, the min-entropy is equal to $H_{\min}(Z) = -\log_2\left[\max_z P(z)\right]$, measured in bits. If the model is deterministic, this maximum is equal to one and the entropy is zero, while for a perfectly random variable the min-entropy achieves its maximum value $\log_2 d$. In our case, the randomness of the outcomes generated by the devices for the pair of inputs x and y reads $H_{xy}(AB) = -\log_2 r_{xy}$, where $r_{xy} = \max_{ab} P(a,b|x,y)$.

All these concepts lay the basis for our first result: a lower bound on the min-entropy of the outcomes produced by two quantum devices violating a Bell's inequality. For a given observed value $\bar{\beta} > \beta_L$ of a Bell inequality, we want to solve the following optimization problem: find the quantum realization, that is, the quantum states and measurements that minimizes the min-entropy of the outcomes. At first sight, solving this minimization problem looks extremely hard, as one should look over all possible quantum states and measurements, in any given space of any dimension, compatible with the observed Bell violation. However, we can tackle this problem using the techniques introduced in Navascues et al. (2007). There, a hierarchy of sets is derived which better and better approximates the set of quantum correlations. The important point is that each of these conditions can be mapped into a semi-definite programming instance, for which there exist efficient numerical techniques. Thus, we can relax the previous optimization problem and solve it over the sets in the hierarchy. Since all of them contain the set of quantum correlations, the obtained solution is a lower bound to the minimum of the min-entropy over quantum correlations (in many cases the lower bound is tight).

Thus, for any violation of any Bell inequality, denoted by $\bar{\beta}$ as above, we can prove that the randomness of a pair of outcomes satisfies

$$H_{xy}(AB) \geq f\left(\bar{\beta}\right), \tag{2.3}$$

where f is a convex function which goes to zero at the point of no violation.

In order to illustrate our results, we plot in Fig. 2.3 the min-entropy for the simplest and best known example of Bell inequality, namely the CHSH inequality. The region below the curve is impossible within quantum physics. As mentioned in the main text, the same bounds can be computed just assuming the validity of the no-signalling principle, and not the entire quantum formalism. The corresponding results are also shown in Fig. 2.3. The derived bounds are worse, as non-signalling correlations are strictly larger than quantum ones.

Appendix B: Device-Independent Quantum Random Number Generator

In this appendix, we exploit (2.3) to construct a novel type of random number generators which are certifiable, private and device-independent. As pointed out by Colbeck in his PhD Thesis (Colbeck 2007), the random character of the generated numbers is guaranteed by the violation of a Bell inequality. As discussed in the main text, we actually propose a randomness expander, a device which expands an initial random seed into a much larger string of random bits.

In order to realize such a randomness expander, we suppose that we have a device, composed of subsystems A and B, that is used n times in succession. The inputs x_i, y_i at each round i are chosen always in the same way and independently of the previous rounds. The amount of randomness needed for this choice can be tuned such that, in the limit of large n, it scales as $\sqrt{n}$. Each use of the device produces outputs a_i and b_i. We denote by $\vec{x} = (x_1, \ldots, x_n)$, and similarly $\vec{y}, \vec{a}$ and $\vec{b}$ the strings of inputs and outputs. Using the previous bound (2.3), we can show that, with probability $1 - \delta$, where δ decreases exponentially with n, the min-entropy of the final string of outputs satisfies

$$H_{\min}\left(\vec{a}, \vec{b}|\vec{x}, \vec{y}\right) \geq n f\left(\tilde{\beta} - \varepsilon\right), \tag{2.4}$$

where ε is a security parameter which can be taken very small and $\tilde{\beta}$ is an estimation of the Bell parameter β derived from the observed symbols. Note that the output randomness scales as n, while the initial randomness needed for the tests scaled as $\sqrt{n}$. Beyond the technical details, which are just sketched here, the importance of this bound comes from the fact that it holds even if the devices have internal memories and can adapt their responses to what was produced in the

previous rounds. This is a significant advance over the device-independent protocols proposed so far and is the crucial feature that makes our protocol practical.

Finally, note that although the output string may not be uniformly random, we are guaranteed that its entropy is bounded by (2.4). The output string can now be classically processed using a randomness extractor (Nisan and Ta-Shma 1999). An extractor is a well-known technique in information theory that with the help of a small private random seed, maps an initial string of bits whose entropy is bounded by k, see (2.4), into k perfect random bits. The use of the extractor, then, concludes the randomness generation (or, more precisely, expansion) process.

Appendix C: Experimental Generation of Private Random Numbers

The experimental realization of our proposal requires the observation of a Bell inequality with the detection loophole closed. This means that almost every event has to be recorded, so that the outputs cannot be deterministically reproduced. This is a strong technological requirement and implies that no photon experiment is possible with current technology, as photon detection is a rather inefficient process. At the moment, the only Bell experiments which are able to close detection loophole consist of trapped atomic particles (Matsukevich et al. 2008, Rowe et al. 2001). Moreover, in our proposal the two devices should also be sufficiently separated so that they do not interact when the measurements are performed. This is needed to assure the tensor product structure in (2.1), which is crucial in the derivation of our results. The only way of guaranteeing this is by considering atoms in two distant traps. At present, the unique setup in the world which satisfies all these requirements is the one in the group of Prof. Monroe, at the University of Maryland. They are able to entangle two atoms in two distant traps and observe a Bell violation with closed detection loophole.

Together with the group of Prof. Monroe, we performed a proof-of-principle demonstration of our proposal. We realize this situation with two ^{171}Yb atomic ions confined in two independent vacuum chambers separated by about 1 m (see Fig. 2.2). First, the atoms are entangled via the coincidence detection of two photons, each one emitted by each ion (Moehring et al. 2007). This process is probabilistic, but when it succeeds, leaves the two ions in a maximally entangled state. The ions are then measured and lead to the violation of the CHSH-Bell inequality. This inequality involves two different measurements per ion. The choice between these two measurements is random and set prior to measurement. Direct interaction between the atoms is negligible and classical microwave and optical fields used to perform measurements on one atom have no influence on the other atom.

To estimate the value of the CHSH inequality $n = 3016$ successive entanglement events were accumulated over the period of about one month. The observed CHSH violation was equal to 2.414 and represents a substantial improvement over previous

results (Matsukevich et al. 2008). Using our theoretical formalism, we can prove that the observed CHSH violation implies that at least $H_{\min}\left(\vec{a},\vec{b}|\vec{x},\vec{y}\right)>42$ new random bits are generated in the experiment with a 99% confidence level. Thus, we can, for the first time, certify that new randomness is produced in an experiment without a detailed model of the devices.

References

Acin, A., Brunner, N., Gisin, N., Massar, S., Pironio, S., & Scarani, V. (2007). Device-independent security of quantum cryptography against collective attacks. *Physical Review Letters, 98*, 230501.

Aspect, A., Dalibard, J., & Roger, G. (1982). Experimental test of Bell's inequalities using timevarying analyzers. *Physical Review Letters, 49*, 1804–1807.

Bell, J. S. (1965). On the Einstein-Podolsky-Rosen paradox. *Physics, 1*, 195–200.

Bell, J. S. (2004). *Speakable and unspeakable in quantum mechanics: Collected papers on quantum philosophy*. Cambridge: Cambridge University Press.

Bennett, C. H. & Brassard, G. (1984). Quantum cryptography: Public key distribution and coin tossing, *Proceedings of the IEEE International Conference on Computers, Systems and Signal Processing*, 175–179.

Bohm, D. (1952a). A suggested interpretation of the Quantum Theory in terms of the "Hidden Variables" I. *Physical Review, 85*, 166–179.

Bohm, D. (1952b). A suggested interpretation of the Quantum Theory in terms of the "Hidden Variables" II. *Physical Review, 85*, 180–193.

Buhrman, H., Cleve, R., Massar, S., & de Wolf, R. (2010). Non locality and communication complexity. *Reviews of Modern Physics, 82*, 665.

Clauser, J. F., Horne, M. A., Shimony, A., & Holt, R. A. (1969). Proposed experiment to test local hidden-variable theories. *Physical Review Letters, 23*, 880–884.

Colbeck, R. (2007). *Quantum and relativistic protocols for secure multi-party computation*, PhD Dissertation, University of Cambridge.

Einstein, A., Podolsky, B., & Rosen, N. (1935). Can quantum-mechanical description of physical reality be considered complete? *Physical Review, 47*, 777–780.

Ekert, A. K. (1991). Quantum cryptography based on Bell's theorem. *Physical Review Letters, 67*, 661–663.

Ekert, A. (2009). Less reality, more security. *Physics World*, 28–32.

Knuth, D. (1981). *The art of computer programming* (2nd ed., Vol. 2). Reading: Addison-Wesley.

Laplace, P. S. (1840). *A philosophical essay on probabilities*, Paris.

Matsukevich, D. N., Maunz, P., Moehring, D. L., Olmschenk, S., & Monroe, C. (2008). Bell inequality violation with two remote atomic qubits. *Physical Review Letters, 100*, 150404.

Moehring, D. L., Maunz, P., Olmschenk, S., Younge, K. C., Matsukevich, D. N., Duan, L.-M., & Monroe, C. (2007). Entanglement of single-atom quantum bits at a distance. *Nature, 449*, 68–71.

Navascues, M., Pironio, S., & Acin, A. (2007). Bounding the set of quantum correlations. *Physical Review Letters, 98*, 010401.

Nielsen, M., & Chuang, I. (2000). *Quantum information and quantum computation*. Cambridge: Cambridge University Press.

Nisan, N., & Ta-Shma, A. (1999). Extracting randomness: A survey and new constructions. *Journal of Computer and System Sciences, 58*, 148–173.

Pironio, S., Acin, A., Brunner, N., Gisin, N., Massar, S., & Scarani, V. (2009). Device-independent quantum key distribution secure against collective attacks. *New Journal of Physics, 11*, 045021.

Pironio, S., Acin, A., Massar, S., Boyer de la Giroday, A., Matsukevitch, D. N., Maunz, P., Olmschenk, S., Hayes, D., Luo, L., Manning, T. A., & Monroe, C. (2010). Random numbers certified by Bell's theorem. *Nature, 464*, 1021–1024.

Rowe, M. A., Kielpinski, D., Meyer, V., Sackett, C. A., Itano, W. M., Monroe, C., & Wineland, D. J. (2001). Experimental violation of a Bell's inequality with efficient detection. *Nature, 409*, 791–794.

Rukhin A., et al. (2008). *A statistical test suite for random and pseudorandom number generators for cryptographic applications*, National Institute of Standards and Technology, Special Publication 800-22 Revision 1, available at http://csrc.nist.gov/publications/PubsSPs.html.

Shor, P. W. (1997). Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Journal on Computing, 26*(5), 1484–1509.

The Marsaglia Random Number CDROM including the Diehard Battery of Tests of Randomness, available at http://www.stat.fsu.edu/pub/diehard/ (2008).

Valentini, A. (2002). Signal-locality in hidden-variables theories. *Physics Letters A, 297*, 273.

Chapter 3
Are There Quantum Effects Coming from Outside Space–Time? Nonlocality, Free Will and "No Many-Worlds"

Nicolas Gisin

Abstract Observing the violation of Bell's inequality tells us something about all possible future theories: they must all predict nonlocal correlations. Hence Nature is nonlocal. After an elementary introduction to nonlocality and a brief review of some recent experiments, I argue that Nature's nonlocality together with the existence of free will is incompatible with the many-worlds view of quantum physics.

Keywords Free will • Nonlocality • Entanglement • "Many-worlds"

3.1 Introduction

Imagine several persons that each separately and independently make choices that have consequences. For the sake of scientific analysis of this banal situation, assume that the same set of persons can repeat again and again the experiment, that is again and again make a free choice and observe its consequence. Moreover, for simplicity, assume that each one has a choice between a finite set of possibilities, that we name inputs, and that the consequences can be catalogued into a finite set of possible outcomes. Once enough data are collected, the probability of the various possible outcomes, given one possible input per person can be estimated. For example, if there are only two persons, that we may name Alice and Bob, and we label their inputs x and y and their outcomes a and b, respectively, the probability reads:$p(a, b \mid x,y)$. For conciseness, we call such a conditional probability distribution, $p(a, b \mid x, y)$, a correlation, see Fig. 3.1.

Correlations are observed every day everywhere and, in particular, in all natural sciences. One could even argue that the scientific activity consists in observing correlations and developing theoretical models that explain them, i.e., that describe

N. Gisin (✉)
Group of Applied Physics, University of Geneva, 1211 Geneva 4, Switzerland
e-mail: Nicolas.Gisin@unige.ch

A. Suarez and P. Adams (eds.), *Is Science Compatible with Free Will?: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience*,
DOI 10.1007/978-1-4614-5212-6_3,

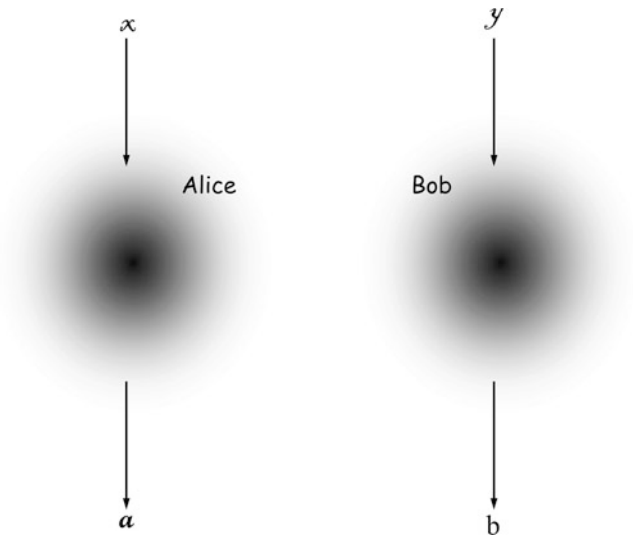


Fig. 3.1 For each run of the experiment, Alice and Bob each freely and independently chose one value x and y, respectively, and input them into their black boxes; the latter then returns one and only one outcome a and b to Alice and Bob, respectively. Note that in order to test condition (3.7), see Sect. 3.2, the experiment has to be repeated many times until the statistics allows one to infer a good approximation of the probability $p(a, b \mid x, y)$

how they happen. For example, if one watches a football or a fieldhockey game on a TV with the sound shut off and one observed that all the players simultaneously stop running, one would speculate, as an explanation based on our implicit theory of the game, that the umpire has whistled.

Surprisingly, the number of categories of explanations for correlations is extremely limited. Before quantum physics, there were only two categories of explanations: Either a first system influences a second one by sending it some information encoded in some physical systems, or the correlated events share some common causes in their common past. For example, in the football or hockey game, all players simultaneously stopped running because in their common past the umpire whistled, i.e., acted as a common cause for all players.

The two categories of explanations are local in the sense that the processes start at a localized place and propagate locally from one place to an adjacent one. Hence, the usual terminology reads *local common cause*, to emphasize the central importance of locality which lies at the core of these explanations

It is difficult to imagine any other sort of explanation. Actually, if one insists that an explanation ought to be a kind of story that plays out in space and time, then I believe there is simply no alternative to the previously mentioned two categories of explanations. Yet, amazingly, quantum physics predicts entirely different kinds of correlations, called nonlocal correlations for reasons described below. Physics has a word for the cause of these nonlocal correlations: entanglement. But physics offers

no story in space and time to explain or describe how these correlations happen. Hence, somehow, nonlocal correlations emerge from outside space–time (for an explanation of this provocative terminology see Sect. 3.5).

3.2 Nonlocal Correlations

Why should anyone believe the existence of nonlocal correlations? Their existence is predicted by quantum theory, as Einstein-Podolsky-Rosen and Schrödinger, Erwin noticed already in 1935 (and actually years before) and also a few other precursors (for a beautiful account of the history of quantum nonlocality read Gilder 2008). But the possibility to directly observe nonlocal correlation seems, at first sight, difficult, if at all possible: indeed, one should observe correlations while simultaneously excluding any explanation of the two categories mentioned in the introduction.

The first type of explanation, i.e., a first system sends information to the second, is quite easy to control, professors do that all the time during exams: they make sure the students can't communicate. In this way professors guarantee that if the exam's results are correlated it is not because one student copied the other, but because they prepared the exam together (i.e., the only remaining explanation is local common cause). In physics, avoiding information exchange is straightforward, at least in principle: separate the correlated events such that nothing propagating at the speed of light can leave a system after the input has been given and reach the other before the outcome has been secured (physicists say that the events are space-like separated). Let us emphasize this point. Bob should observe his outcome, i.e., the consequence of his choice, before anything propagating at most at the speed of light could reach him carrying any information about Alice's choice; and vice versa Alice should observe her outcome before any influence of Bob's choice, propagating at the speed of light, could reach her. Experiments that do not strictly fulfill this condition are said to suffer from the "locality loophole."

But what about the second category of explanation, how could one experimentally rule out any local common cause explanation? The finding of a solution to this problem is John Bell's main contribution to physics (Bell 1964). It is pretty easy to formalize; let's have a look at Fig. 3.1. Alice and Bob should each have access to only a limited part of space and time. In particular one should be able to bound where and when the input choices are made (one by Alice, another one by Bob), and bound where and when the outcomes are produced and registered. Note that the inputs and outcomes are standard (i.e., classical) variables: they can be copied, remembered, and processed as any of the usual information we confront daily. For concreteness, assume Alice and Bob put their inputs and outcomes on the internet so that, after some time, everyone can access them. Let's assume that the correlation $p(a, b \mid x, y)$ has a common cause explanation. Let λ denote this common cause. We do not need to know what λ is, so far it is just a symbol. We make only two assumptions about λ, a serious one and a technical one. First the

serious one: we assume that λ doesn't contain any information about Alice's and Bob's free choice: the inputs x and y are independent of λ. Note that this excludes hyper-determinism: Alice and Bob can make truly free choices (I'll come back to this). This assumption can be formalized: $p(x) = p(x \mid \lambda)$. Or equivalently: $I(x : \lambda) = 0$: the (Shannon) mutual information between x and λ is nil. The second assumption, the technical one, guarantees that one can "count" and "weight" all the possible common causes λ.[1] A priori one doesn't know λ, but all that is necessary is to be able to associate (possibly unknown) probability weights with all the possible λ's. For example, it suffices to assume that there are only countably many possible common causes, possibly infinitely countable (as the integers). Or, if one insists on the possibility of a continuous infinity of common causes (e.g., the inputs depend on the temperature of some location in their common past), then one has to assume that the set of λ's is equipped with a measure such that one can integrate over the space of λ's (for an example showing that this assumption is necessary see Pitowsky 1982).

Now, if the correlation $p(a, b \mid x, y)$ has some local common cause explanation that satisfies the two above mentioned assumptions, then, for any given λ, the two events are independent:

$$p(a, b|x, y, \lambda) = p(a|x, \lambda) \cdot p(b|y, \lambda) \tag{3.1}$$

Since, a priori, one doesn't know λ one has to attribute a certain probability to each of them: denote $\rho(\lambda)$ the probability that the actual common cause is λ. Note that the function $\rho(\lambda)$ may be unknown, but it is part of the local common cause category of explanations to assume that a ρ exists. Consequently, any common cause explanation of correlations takes the form:

$$p(a, b|x, y) = \sum_{\lambda} \rho(\lambda)\ p(a|x, \lambda) \cdot p(b|y, \lambda) \tag{3.2}$$

or if a continuous infinity of λ's is assumed:

$$p(a, b|x, y) = \int_{\lambda} \rho(\lambda)\ p(a|x, \lambda) \cdot p(b|y, \lambda)\ \mathrm{d}\lambda \tag{3.3}$$

Agreed? The rest of the argument is elementary mathematics. In brief, not all correlations $p(a, b \mid x, y)$ can be put in the form (3.2) or (3.3). Hence, if one observed a correlation that can't be written as (3.2) or (3.3), one has observed a correlation that can't be explained by local common causes. John Bell introduced a simple inequality, now generalized to entire families of so-called Bell inequalities, that are necessarily satisfied by all correlations of the form (3.2) or (3.3)

[1] Note that one can group the λ's into equivalence classes where two λ's are said equivalent iff they determine precisely the same probabilities $p(a, b \mid x, y, \lambda)$; hence it suffices that one can "count" the equivalence classes.

(Bell 1964). We'll soon see an example: (3.6) and (3.7). Hence a violation of a Bell inequality is the signature of a correlation that can't be written as (3.2) or (3.3).

At this point it is worth emphasizing the interpretation of λ. Historically the λ were thought of as local hidden variables by physicists whose hope was to restore some sort of local classical physics. A more modern view consists in viewing λ as the physical state of the systems as described by any possible future theory. Hence, the violation of a Bell inequality tells us something not only about today's quantum physics but also about any possible future theory compatible with today's experiments. That today's experiments tell us something important about any possible future theory is a rare and remarkable fact! Note furthermore how unrestricted λ is: it could be the state of the entire Universe, except that λ can't determine Alice and Bob's input choices x and y. In this sense it is not λ that is especially local, all that is assumed local is that Alice's system is not influenced by Bob's distant choice and vice versa that Bob's system is not influenced by Alice's choice.

As a simple example of a correlation that can't be explained by common causes, consider the case where Alice and Bob have only to carry out a binary choice that we label 0 and 1, i.e., $x, y \in \{0, 1\}$, and their outcomes are also binary: $a, b \in \{0, 1\}$. Note that this is the simplest possible case: with fewer inputs there would be no choice at all and with fewer outcomes the choices would have no consequences. The example goes as follows. Alice's outcome a is random: $p(a|x) = \frac{1}{2}$ for all inputs x; similarly Bob's outcome is random: $p(b|y) = \frac{1}{2}$ for all inputs y. But the two outcomes are correlated: whenever it so happens that Alice and Bob both made the choice 1, i.e., $x = y = 1$, then their outcomes always differ: $p(a \neq b|x = y = 1) = 1$, and for all other combinations of input choices the outcomes are always equal. Since $x = y = 1$ if and only if $x \cdot y = 1$, this simple correlation can be captured with a simple relation:

$$x \cdot y = 0 \Rightarrow a = b \tag{3.4}$$

$$x \cdot y = 1 \Rightarrow a \neq b \tag{3.5}$$

Note that this relation can be cast into a simple equation $a + b = x \cdot y$ (addition modulo 2), hence nonlocality shouldn't be hidden behind complex mathematics: the concepts are complex, not the maths. Let's analyze this correlation and look for a local common cause explanation. For this purpose we consider the following figure of merit:

$$\begin{aligned} S = {} & p(a = b|x = 0, y = 0) + p(a = b|x = 0, y = 1) \\ & + p(a = b|x = 1, y = 0) + p(a \neq b|x = 1, y = 1) \end{aligned} \tag{3.6}$$

Any local common cause λ should, for all possible choices x by Alice define an output a (or define a probability for the outcome a), and similarly for Bob. For instance, one of the possible λ is such that $a = b = 0$ whatever the inputs. For such

a λ our figure of merit S takes the value 3: the first 3 terms in (3.6) take value 1, but the last one is 0. It is not difficult to analyze all possible deterministic λ (those λ's that determine one and only one outcome on each side for any possible inputs), indeed there are only $2^2 \cdot 2^2 = 16$ such λ's. Analyzing these 16 λ's one can easily convince oneself that our figure of merit S never reaches a value larger than 3. And nondeterministic λ's will not perform better (note that they can always be analyzed as statistical mixtures of the 16 deterministic λ's). Consequently, all correlations explainable by local common causes satisfying the following inequality, named a Bell inequality:

$$S \leq 3 \tag{3.7}$$

Let me note for the more specialized readers that this inequality is strictly equivalent to the well-known CHSH-Bell inequality: it suffices to note that the usual $E(x, y) \equiv p(a = b|x, y) - p(a \neq b|x, y)$ can equally be written as $E(x, y) = 2p(a = b|x, y) - 1 = 1 - 2p(a \neq b|x, y)$, the common form of the CHSH-Bell inequality follows then from (3.6) and (3.7): $E(0, 0) + E(0, 1) + E(1, 0) - E(1, 1) \leq 2$.

To conclude this section, let us emphasize the main point: the two categories of local explanations for correlations can be experimentally tested. For this purpose one should observe correlations that violate some Bell inequality, as for example (3.7), while making sure that the two observers, Alice and Bob, can't be influenced by any signal coming from the other side propagating at the speed of light (or slower). If such correlations are observed, there is no choice but to admit that there are correlations that can't be explained by any story in space and time. Such correlations are thus said to be nonlocal: there is no "local explanation," that is no explanation based on local causes that propagate from one place to adjacent ones.

3.3 Experimental Nonlocality

In this section I review some of the recent experiments, though without any of the important technicalities. Already in the famous Alain Aspect experiment of 1982 the sides were space-like separated and the inputs chosen at "random" at the last moment so that no light-signal could explain the observed correlation (Aspect et al. 1982). Admittedly there were no human Alice and Bob making free choices, only some pseudo-random, even somewhat periodic, choices were made by appropriate electronics. For scientists this was already extremely convincing, though since that time better experiments definitively closing the locality loophole have been performed (Gisin and Zbinden 1999, Tittel et al. 1999, Weihs et al. 1998).

All the above experiments observed correlations that violate a Bell inequality (3.7). However, there is a little catch: in all experiments with photons (particles of light) there is often no outcome at all. For example, Alice inputs her choice, but nothing happens. Physicists understand why this is so, the photon got lost somewhere, or the detector supposed to register the tiny bit of energy carried by a single photon failed to

do so (no real detector has 100% efficiency), etc. Nevertheless, this is a serious loophole, called the detection loophole. Indeed, it could well be that the detection probability is influenced by the local common cause λ. Today, two experiments have closed this loophole using not photons, but ions (Matsukevich et al. 2008, Rowe et al. 2001). This was a necessary step; however, in those two experiments the distance between Alice and Bob was insufficient to close the locality loophole. Hence, an experiment closing simultaneously the detection and the locality loophole is still awaited. Almost no physicist expects a surprise, certainly I do not expect any surprise, but the logical possibility remains and ought to be closed by further experiments.

So are we at the end? Do we have to conclude that Nature is nonlocal? Are there really correlations that can't be explained within space–time, i.e., that somehow emerge from outside space–time? The situation clearly deserves further scrutiny. In the remainder of this section I would like to analyze two local explanations together with experimental tests.

The first explanation is, I believe, very intuitive. Everything looks as if the two parties somehow communicate behind the scene (Bell 1993); hence, since they can't communicate at a speed equal to or slower than the speed of light, let's assume they do so at a speed faster than light. Such an assumption doesn't respect the spirit of Albert Einstein's relativity, but this wouldn't be the first time that an accepted theory has to be revised (Gisin 2005). Moreover, it is not crystal clear that such "communication behind the scene" would contradict relativity; indeed, one could imagine that this communication remains for ever hidden to humans, i.e., that it could not be controlled by humans, only Nature exploits it to produce correlations that can't be explained by usual common causes. To define faster than light hidden communication requires a universal privileged reference frame in which this faster than light speed is defined. Again, such a universal privileged frame is not in the spirit of relativity, and also clearly not in contradiction: for example the reference frame in which the cosmic microwave background radiation is isotropic defines such a privileged frame. Hence, a priori, a hidden communication explanation is not more surprising than nonlocality. It also has the very nice feature that it can be experimentally tested. The idea is to perform the measurements on both sides, i.e., give the inputs and collect the outcomes, quasi-simultaneously. Hence, Bob's outcome can't be influenced by any hypothetical hidden communication and vice versa for Alice's outcome. If the observed correlation is still nonlocal, i.e., still violates Bell's inequality, then either the hypothesis of hidden communication is ruled out, or the speed of the hidden communication is faster than the bound set by the experimental condition, in particular by the accuracy of the synchronous timing and by the distance separating Alice and Bob. But there remains a conceptual difficulty: since we do not know which is the privileged reference frame, we do not know in which reference frame the event should be simultaneous. Philippe Eberhard suggested exploiting the rotation of Earth around its axes to scan all possible reference frames in 12 h. This experiment has been carried out recently near Geneva (Salart et al. 2008a) and has set very stringent bounds on the speed of any hypothetical hidden communication: more than 10,000 or 100,000 times the speed of light, depending on technical details (see also the recent paper Cocciaro et al. 2011). While finishing this contribution, with colleagues we found a general argument against finite but supraluminal hidden influences, see Bancal et al. (2012).

Before we come to the second alternative, let me mention that there is another way to define the faster than light hypothetical hidden communication: it could be that it is the inertial reference frame of the observer that determines that privileged frame. This very interesting idea was put forward by Suarez and Scarani (1997). A consequence of this assumption is that, thanks to relativity, if the two observers Alice and Bob move apart fast enough, they could both, each in its own inertial reference frame, perform the measurement before the other, a so called before–before situation. This experiment was also be carried out in Geneva (Gisin et al. 2000; Stefanov et al. 2002; Suarez 2001; Zbinden et al. 2001), and the observed correlation was still nonlocal: the proposal by Suarez and Scarani could be falsified.

The second way out of the conclusion "Nature is nonlocal" speculates on the fact that in actual experiments it is not so easy to determine when a choice is made and when an outcome is produced. Ideally, human Alice and Bob should make conscious choices, but in all experiments so far the choices are delegated to random number generators (or, even, no active choice is made, one merely argues—quite convincingly in my opinion—that the measurement settings are unknown to λ and to the particles until the moment they reach the measurement apparatuses). Delegating the choices to random number generators is pretty fine with me. After all, all that is required is that the choices are independent of the common past. Assuming that Alice and Bob's common past drives all choices made locally at Alice and Bob's locations by appropriate electronic or quantum devices seems to imply some sort of hyper-determinism that would make all Science an illusion (one could never decide to make an experiment, hence one could not test theories). Accordingly, let's concentrate on the idea that the outcome might, in fact, be determined much later than usually thought (Franson 1985). For example, two physicists, Lajos Diosi and Roger Penrose, independently proposed that an outcome is produced only once a mass has moved significantly (both proposed precise formulas relating the time of the outcome and the motion of the mass, their formulas agree within a factor 2, see e.g. Adler 2007). The motivation for this proposal lies in the difficulty of combining general relativity and quantum physics. But, never mind, here it suffices to note that in usual experiments the outcomes are collected in a computer's memory, hence without motion of any significant mass (electrons are very light). Hence, all observed violations of the Bell inequality could be explained by slower than light influences: the influence has plenty of time to arrive before any mass moves significantly (Kent 2009). Fortunately, once again, this assumption of delayed outcomes can be experimentally tested. We coupled our detectors to a piezzo that could push a mirror and could thus falsify the Diosi-Penrose explanation of correlations violating Bell inequality (Salart et al. 2008b).

No doubt further assumptions will appear. However, the huge amount of experimental data and the enormous predictive power of quantum physics very convincingly supports the view that Nature is nonlocal. So, how do physicists incorporate this amazing conclusion in their world view? Well, most simply don't care, most don't realize that they are living at a time of a huge conceptual revolution; sadly, most physicists would not have recognized Nicolaus Copernicus nor Galileo,

Galilei had they been contemporaries of these giants that carried out huge conceptual revolutions. But there are exceptions that one may classify, roughly, into two categories: the many-worlds lovers and the others (to which I belong).

3.4 Against Many-Worlds

Basically the solution proposed by the many-worlds view of quantum physics, also called the multiverse, is to deny that experiments have unique outcomes (for a long list of various versions of the many-world view see Kent 2010). According to this view, everything is quantum, once and for ever. Hence, the entire reasoning of Sect. 3.2 collapses: there are no inputs and no outputs! Actually, the motivation for many-worlds is not nonlocality, but the fact that today's quantum theory offers no answer as to when a quantum measurement is finished. Hence, they conclude: quantum measurements are never finished, everything gets into an enormously complex state of superposition. Somehow, the only real thing is the Hilbert space and the linearity of Schrödinger's equation.[2]

I won't try to present the many-worlds view any further; from the little above it should already be clear that I am not sympathetic with this view. But why am I so dismissive with this view while, at the same time, very open to all sort of assumptions like those presented in the previous section? Two reasons. First, all the assumptions presented in the previous section have an explanatory power. Moreover they could even be experimentally tested (and—even better for me—using technologies available in my lab!). On the contrary, I do not see any explanatory power in the many worlds view: it seems to be made just to prevent one from asking (possibly provocative) questions. Moreover, it has built into it the impossibility of any test: all its predictions are identical to those of quantum theory. For me, it looks like a "cushion for laziness" (*un coussin de paresse* in French).

And there is a second, decisive, reason to reject the many-worlds view: it leaves no space for free will. I know that I enjoy free will much more than I know anything about physics. Hence, physics will never be able to convince me that free will is an illusion. Quite contrary, any physical hypothesis incompatible with free will is falsified by the most profound experience I have about free will.

So, would I have rejected Newtonian classical mechanics had I lived before quantum physics? Probably not. Indeed, classical physics leaves open the possibility that free will can somehow interface with the deterministic Newtonian equations: free will could set up some potential that could slightly influence particles' motion. This would be something like René Descartes' pineal gland. In standard quantum

[2] Years ago, I once argued that the many-worlds doesn't seem compatible with Occam's razor principle (Gisin and Percival 1993). As answer I got the following: "Occam's razor should not be applied to the physical world, but be applied to the Schrödinger equation; don't add any term to this beautiful equation" (Zeh 1993). The linearity of the Schrödinger equation was assumed more real than our physical universe!

physics such an interface between free will and physics could be even simpler: free will could influence the probabilities of quantum events. This is, admittedly, a vague and not very original idea; but the important thing is that there is no obvious definite contradiction between free will and standard quantum physics. However, the situation with the many-worlds view is very different.

In the many-worlds view all possibilities coexist on an equal footing. Accordingly, a being enjoying free will can't merely interact with one state of affair of the physical world, but has to interact with the enormous superposition of all possible states of affairs! But, most likely, if a specific interaction with one possible state of affairs produces a desired effect, this very same specific interaction with most of the other—equally real according to many-worlds—state of affairs would produce uncontrolled random effects. Hence, it seems that there is no way to maintain a possible window for free will in the many-worlds view. Consequently, I believe the many-worlds view is excluded by our daily experience.

A possible way out of the above reasoning could be to envisage that the being enjoying free will is also in an enormous superposition state, and that the branches of this superposition match the branches of the superposition of the physical world. Hence, in each branch a story similar to the one sketched above in the case of Newtonian classical mechanics could hold (to maintain hope). But this "way out" is an illusion. Indeed, it would imply that the being enjoying free will actually never makes one and only one decision, nor experience one and only one consequence of his choice: he would make a superposition of all choices and experience all possible consequences. In brief, such a being would enjoy no free will at all.

In summary, superpositions and entanglement forever, i.e., the many-worlds "solution" to nonlocality, is not compatible with our most intimate experience as beings who enjoy free will. I make choices that have consequences; hence superpositions and entanglement must end somewhere. And the fact that today's physics doesn't know where they stop doesn't affect this conclusion at all.

3.5 What Could It Mean that Nonlocal Correlations Emerge from Outside Space–Time?

In physics we develop mathematical models that allow us to compute the outcome of some experiments, or their outcome probabilities, i.e., we have equations. However, this is only half of theoretical physics. We also develop stories that describe how things happen. One story, for example, goes as follows: "The moon attracts the water in the ocean, hence produces the tides." Or, in another example, we describe the relation between temperature and pressure of a gas by a story like: "The gas is made out of trillions of little particles that move in all directions; the warmer the gas the faster the particles on average; when the particles hit the recipient containing the gas they exercise a small force on it, hence the trillions of particles all together exercise some pressure on the container; finally the pressure is larger when the

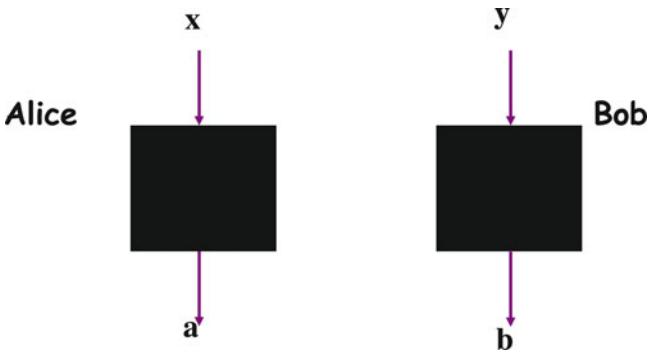


Fig. 3.2 Example of a possible new tool to talk about nonlocal correlations. The inputs x and y and the outcomes a and b are all bits. As soon as an input is fed into a box, the box produces a random outcome. But the outcomes hold the following promise: $a + b = xy$ (addition modulo 2). This promise holds irrespective to the inputs time ordering and independently of the distance between Alice and Bob. As explained in the text, this is a simple and powerful example of a nonlocal correlation. In quantum physics the relation is not exactly satisfied, but the measurement outcomes a and b tend to "attract" each other so as to satisfy $a + b = xy$ in about 85% of the cases

average velocity of the particles is larger." Who has ever started a physics course with equations and not with a story? Clearly, in physics we need stories as much as equations. For this purpose we have a catalogue of possible tools to tell our stories. Until recently, all stories took place in space–time. But, this story-toolbox evolves as our theories evolve in parallel with our mathematics-toolbox; see for example from the story-toolbox used today to talk about the deformation of space–time in general relativity (typically a weight that deforms a two-dimensional sheet).

However, as we have seen in Sect. 3.2 no story in space–time can describe nonlocal correlations: we have no tool in our present day story-toolbox to talk about nonlocal correlations. Hence, we usually say things like "event A influences event B," or "event A has a spooky action at a distance on event B" or "event A causes a collapse of the wavefunction at location B." But we know that this is all wrong: there is no time ordering between the events A and B; hence no story in time is appropriate. Moreover, the distance between A and B is irrelevant; hence the distance should not occur in our story. The usual reaction to this situation is to give up the search for any story, that is, to give up the very possibility to make sense of nonlocal correlations, i.e., to understand them. Some physicists simply claim that the maths are too complicated; hence, they claim, we can't complement the equations by good stories. But we have seen that the maths are trivial: this can't be an excuse to give up!

Admittedly we need to enlarge our story-toolbox. A difficulty is that the new tool must include some strange features that can't be described within space–time. I am confident that with future quantum technologies this new piece in our story-toolbox will be familiar to future generations. Let me give an example of how this new piece could look. Imagine a pair of boxes, see Fig. 3.2. Each box can be fed by an input,

denoted x for the first box and y for the second; as soon as a box receives an input, it produces an outcome denoted a and b for the first and second box, respectively. For simplicity imagine that the inputs and outcomes are binary: both inputs and both outcomes are simply bits, i.e., a "0" or a "1." Assume furthermore that the outcomes of each box are random: each box just produces noise. But now, in order to tell a story about entangled boxes, assume that the outcomes of the two boxes tend to attract each other in such a way as to satisfy the following promises: $a + b = xy$ (addition modulo 2). This is identical to the relations (3.4) and (3.5). This new tool is unfamiliar to us, but it is quite simple. Moreover, it contains the require essence to tell stories about nonlocal correlations: the promise $a + b = xy$ holds irrespective of the input timing and holds independently of the distance between the two boxes; furthermore the correlation $p(a, b \mid x, y)$ is nonlocal (i.e., it can't be described by local common causes because it violates the Bell inequality (3.7)). This tool is well known to specialists and is referred to as "nonlocal boxes" (sometimes also called a PR-box according to its inventor Popescu and Rohrlich 1994). It shares with quantum nonlocal correlations, the important feature that it can't be cloned (Massanes et al. 2006) (and the proof is very simple: again a nice story), accordingly one can also use this tool to tell simply about quantum cryptography (Acin et al. 2006). Finally, let me mention that with such a nonlocal box all quantum correlations corresponding to two maximally entangled qubits can be reproduced (Cerf et al. 2005), hence the nonlocal box contains enough nonlocality to encompass the most usual correlations one encounters in quantum physics. However, to be fair, I should add that this new tool is insufficient to tell a story describing quantum teleportation (Short et al. 2006).

Hence, more tools are needed in our story-toolbox. Looking for such new tools, however, is not standard research in physics. Nonetheless, can we expect physics to make progress and be appreciated by the public, as it should, if we give up the possibility to tell stories about it?

3.6 Conclusion

We have seen that any proper violation of a Bell inequality implies that all possible future theories have to predict nonlocal correlations. In this sense it is Nature herself that is nonlocal (Sect. 3.2). But how can that be? How does Nature perform the trick (Gisin 2009a)? Leaving aside some technical loopholes, like a combination of detection and locality loopholes, the obvious answer, already suggested by Bell (1993), is that there is some hidden communication going on behind the scene. A first meaning of "behind the scene" could be "beyond today's physics," in particular beyond the speed limit set by relativity. We have seen how this interesting idea can be experimentally tested (Sect. 3.3) and how difficult it is to combine this idea with no-signaling (Appendix 3.7 and Bancal et al. 2012).

Hence, it is time to take seriously the idea that Nature is able to produce nonlocal correlations.[3] There are several ways of formulating this:

1. Somehow God plays dice with nonlocal die: a random event can manifest itself at several locations.
2. Nonlocal correlations merely happen, somehow from outside space–time (from), in the sense that no story in space–time can describe how they happen (see Sect. 3.5).
3. The communication behind the scene happens outside space–time.
4. Reality happens in configuration space, what we observe is only its shadow in three-dimensional space (this might be the closest to the description provided by standard quantum physics).[4]

Admittedly, the situation is serious, so much so that despite the vast evidence further scrutinies should be undertaken. However, at this point we should have the courage to also seriously consider the possibility that Nature is indeed truly and deeply nonlocal.[5]

At this point one should ask oneself whether this is really new or whether similar conclusions already follow from the nondeterministic characteristic of quantum physics? Indeed, one could argue that nondeterminism implies that the cause originates from elsewhere, i.e., somewhere outside space–time. But this doesn't sound very convincing. I have no problem with the idea that certain objects may have an intrinsic propensity to spontaneously act in a stochastic manner. Furthermore, stochasticity by itself could act purely locally. Hence, with nonlocality we face something deeply different.

One logical possibility to avoid the entire argument—and hence the conclusion "Nature is nonlocal"—is to deny the possibility to freely choose inputs and/or collect measurement outcomes. One could invoke some hyperdeterminism such that the state of the universe λ necessarily determines the inputs x and y, but this seems to me like giving up the entire scientific enterprize. Indeed, with such a totalitarian determinism there would be no way to test one's scientific theories. Alternatively one could deny that measurements have outcomes, or at least that it takes in fact much longer for an outcome to be definitive than usually thought. An example, discussed in Sect. 3.3, could be that a measurement outcome is definitive only once a mass has significantly moved. This interesting explanation of the observed correlation could be experimentally falsified. Another example could be that a measurement is finished only once a human becomes conscious of its outcome . . . but then, as Bell, John put it, "does that human need to have a PhD?". Clearly such ideas are ill defined, though they deserve further scrutiny. Finally, pushed to the extreme, one

[3] Some conclude that it must be realism that is faulty. But I don't see in which sense this could save locality? Moreover, realism is often confused with determinism, an uninteresting terminology issue (see Gisin 2012).

[4] Talk delivered at the first John Stewart Bell prize award ceremony (Gisin 2009b).

[5] Many physicist hate this conclusion because they fear that it allow faster than light signaling. Hence, let me emphasize that nonlocality does not necessarily imply faster than light signaling. Actually, today's paradigm for most specialists is *nonlocality without signaling*.

could argue with the many-worlds "lovers" that measurements don't have outcomes, that all possible outcomes remain potential in some huge superposition state containing all possibilities on an equal footing. I have argued in Sect. 3.4 that such an extreme view is uninteresting and necessarily false because it is incompatible with free will. Admittedly, no physical theory so far has ever been able to include free will in an interesting way; however, the many-worlds view seems to be the first one totally incompatible with our most intimate experience of free will.

3.7 Appendix: Hidden Communication Without Hidden Variables

Experiments can only set bounds on the speed of any possible faster than light hidden communication. What about infinite speed? And could a theoretical argument refute the possibility of hidden communication at an arbitrarily fast but finite speed?

Let me first briefly comment on the idea of hidden communication at infinite speed. Frankly, I have difficulties making sense of such an assumption: essentially it implies that everything could instantaneously influence everything else (Garisto 2002).

Interestingly, however, the second question has at least a partial and positive answer. Indeed, one can prove that there are 3-party scenarios in which any explanation of distant correlations based purely on hidden communication (at any finite speed), hence without any additional local variable λ, would allow one to signal faster than light (Ryff 2009, Scarani and Gisin 2005). The argument runs as follows (Gisin 2009b). Imagine that the three players, Alice, Bob and Charlie, share a GHZ state of three qubits: $(|000\rangle + |111\rangle)/\sqrt{2}$. Alice is far both from Bob and from Charlie. Bob and Charlie are not as far from each other, but still far enough that their input–outcome events are space-like separated, see Fig. 3.3. Further, imagine that Bob and Charlie synchronize their events so well that there is no time for the hidden communication to influence each other. Consequently, if Alice does nothing, but Bob and Charlie measure their qubits in the standard $\{|\,0\,\rangle, |\,1\,\rangle\}$ basis, then they observe random and uncorrelated outcomes. Indeed, all qubits are locally in a random state and there is, by assumption, no time for any influence (even at a speed possibly faster than light, but finite). If, however, Alice makes a

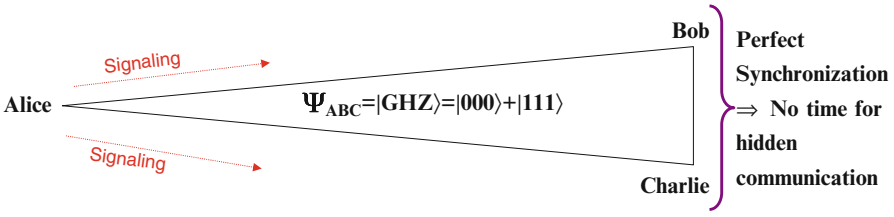


Fig. 3.3 In such a configuration, if all correlations are due to hidden communication behind the scene, then Alice can signal faster than light to Bob/Charlie

measurement, also in the standard basis, long enough before Bob and Charlie (in the privileged reference frame) so that the hidden communication from Alice to Bob and to Charlie has time to arrive, then Bob and Charlie's outcomes are correlated: they are both equal to Alice's outcome. Hence, if Bob and Charlie compare their results, they know whether Alice made a measurement or not, i.e., there is signaling from Alice to (Bob, Charlie). Note that comparing Bob and Charlie's result takes some time, but since Alice could be arbitrarily far away, there is clearly a possibility that the signaling from Alice to (Bob, Charlie) is faster than light.

The above argument illustrates how difficult it is to modify quantum physics while maintaining nonlocality without signaling. However, the sketched argument is clearly of limited scope: it is easy to avoid signaling by adding some local variables λ and by assuming that if the hidden communication doesn't arrive on time, then the outcomes are determined by these additional λ's. It is thus desirable to extend the argument to include mixed models, that is a mix of hidden communication and additional local variables λ. It would be nice to show that any such mixed model necessarily activates signaling in some multipartite scenarios. I find this research program highly interesting.

While finishing writing this contribution a general answer to this question has been found (Bancal et al. 2012). No combination of hidden local variables and hidden communication at finite speed can satisfy both:

1. Reproduce the quantum predictions whenever the hidden communication arrives on time.
2. Remains nonsignalling at the level of measurement inputs and outcomes, i.e., at the human level.

References

Acin, A., Gisin, N., Masanes, L.I. (2006). From Bell's theorem to secure quantum key distribution. *Physical Review Letters*, *97*, 12040.

Adler, S. (2007). Collapse models with non-white noises. *Journal of Physics A*, *40*, 755.

Aspect, A., Grangier, P., Roger, G. (1982). Experimental realization of Einstein-Podolsky-Rosen-Bohm Gedankenexperiment: A new violation of Bell's inequalities. *Physical Review Letters*, *49*, 91–94.

Bancal, J.D., Pironio, S., Acin, A., Liang, Y.C., Scarani, V., Gisin, N. (2012). Quantum Nonlocality Based on Finite-Speed Causal Influences Leads to Superluminal Signaling. *Nature Physics*, in press 2012, http://arxiv.org/pdf/1110.3795.pdf.

Bell, J.S. (1964). On the Einstein-Podolsky-Rosen paradox. *Physics*, *1*, 195–200; reprinted in: Bell, J.D. (1987). *Speakable and unspeakable in quantum mechanics: collected papers on quantum philosophy*. Cambridge: Cambridge University Press, revised edition 2004.

Bell, J.S. (1993). Interview. In P.C.W. Davies, & J.R. Brown (Eds.), *The ghost in the atom* (pp. 45–57). Cambridge: Cambridge University Press.

Cerf, N.J., Gisin, N., Massar, S., Popescu, S. (2005). Simulating maximal quantum entanglement without communication. *Physical Review Letters*, *94*, 220403.

Cocciaro, B., Faetti, S., Fronzoni, L. (2011). A lower bound for the velocity of quantum communications in the preferred frame. *Physics Letters A*, *375*, 379–384.

Franson, J.D. (1985). Bell's theorem and delayed determinism. *Physical Review D*, *31*, 2529–2532.
Garisto, R. (2002). What is the Speed of Quantum Information? http://arxiv.org/pdf/quant-ph/0212078v1.pdf
Gilder, L. (2008). *The age of entanglement*. New York: Alfred A. Knopf.
Gisin, N. (2005). Can Relativity be Considered Complete? From Newtonian Nonlocality to Quantum Nonlocality and Beyond. http://arxiv.org/pdf/quant-ph/0512168v1.pdf
Gisin, N. (2009a). Quantum nonlocality: How does nature do it? *Science*, *326*, 1357–1358.
Gisin, N. (2009b). http://www.gapoptique.unige.ch/wiki/news:20090816-0000-pngatfbp
Gisin, N. (2012). Non-realism: Deep thought or a soft option? *Foundations of Physics*, *42*, 80–85.
Gisin, N., & Percival, I.C. (1993). Stochastic wave equations versus parallel world components. *Physics Letter A*, *175*, 144–145.
Gisin, N., Scarani, V., Tittel, W., Zbinden, H. (2000). Optical tests of quantum nonlocality: from EPR-Bell tests towards experiments with moving observers. *Annals of Physics*, *9*, 831–841.
Gisin, N., & Zbinden, H. (1999). Bell inequality and the locality loophole: Active versus passive switches. *Physics Letters A*, *264*, 103–107.
Kent, A. (2009). A proposed test of the local causality of spacetime. In *Quantum reality, relativistic causality, and closing the epistemic circle: Essays in honour of Abner Shimony* (pp. 369–378). Berlin: Springer.
Kent, A. (2010). One world versus many: the inadequacy of Everettian accounts of evolution, probability, and scientific confirmation. In S. Saunders, J. Barrett, A. Kent, D. Wallace (Eds.), *Many worlds? Everett, quantum theory and reality*. Oxford: Oxford University Press.
Massanes, L.I., Acin, A., Gisin, N. (2006). General properties of nonsignaling theories. *Physical Review A*, *73*, 012112.
Matsukevich, D.N., et al. (2008). Bell inequality violation with two remote atomic qubits. *Physical Review Letters*, *100*, 150404.
Pitowsky, I. (1982). Resolution of the EPR and Bell paradoxes. *Physical Review Letters*, *48*, 1299–1302.
Popescu, S., & Rohrlich, D. (1994). Nonlocality as an axiom. *Foundations of Physics*, *24*, 379–385.
Rowe, M.A., et al. (2001). Experimental violation of a Bell's inequality with efficient detection. *Nature*, *409*, 791–794.
Ryff, L.C. (2009). Bell's Conjecture and Faster-Than-Light Communication. http://arxiv.org/pdf/0903.1076v2.pdf
Salart, D., Baas, A., Branciard, C., Gisin, N., Zbinden, H. (2008a). Testing the speed of 'spooky action at a distance'. *Nature*, *454*, 861–864.
Salart, D., Baas, A., van Houwelingen, J.A.W., Gisin, N., Zbinden, H. (2008b). Spacelike separation in a bell test assuming gravitationally induced collapses. *Physical Review Letters*, *100*, 220404.
Scarani, S., & Gisin, N. (2005). Superluminal hidden communication as the underlying mechanism for quantum correlations: Constraining models. *Brazilian Journal of Physics*, *35*, 328–332
Short, A.J., Popescu, S., Gisin, N. (2006). Entanglement swapping for generalized nonlocal correlations. *Physical Review A*, *73*, 012101.
Stefanov, A., Zbinden, H., Gisin, N., Suarez, A. (2002). Quantum correlations with spacelike separated beam splitters in motion: Experimental test of multisimultaneity. *Physical Review Letters*, *88*, 120404.
Suarez, A. (2001). Is There a Real Time Ordering Behind the Nonlocal Correlations? http://uk.arxiv.org/pdf/quant-ph/0110124v1
Suarez, A., & Scarani, V. (1997). Does entanglement depend on the timing of the impacts at the beam-splitters? *Physics Letters A*, *232*, 9–14.
Tittel, W., Brendel, J., Zbinden, H., Gisin, N. (1998). Violation of bell inequalities by photons more than 10 km apart. *Physical Review Letters*, *81*, 3563–3566; ibid (1999) *Physics Review A*, *59*, 4150–4163.

Weihs, G., Jennewein, T., Simon, C., Weinfurter, H., Zeilinger, A. (1998). Violation of Bell's inequality under strict Einstein locality conditions. *Physical Review Letters*, *81*, 5039–5043.

Zbinden, H., Brendel, J., Gisin, N., Tittel, W. (2001). Experimental test of nonlocal quantum correlation in relativistic configurations. *Physical Review A*, *63*, 022111.

Zeh, H.D. (1993). There are no quantum jumps, nor are there particles! *Physics Letters A*, *172*, 189–192.

Chapter 4
Can Free Will Emerge from Determinism in Quantum Theory?

Gilles Brassard and Paul Raymond-Robichaud

What is proved by impossibility proofs is lack of imagination
John Bell

Imagination is more important than knowledge
Albert Einstein

Abstract Quantum mechanics is generally considered to be *the* ultimate theory capable of explaining the emergence of randomness by virtue of the quantum *measurement* process. Therefore, quantum mechanics can be thought of as God's wonderfully imaginative solution to the problem of providing His creatures with free will in an otherwise well-ordered Universe. Indeed, how could we dream of free will in the purely deterministic Universe envisioned by Laplace if everything ever to happen is predetermined by (and in principle calculable from) the actual conditions or even those existing at the time of the Big Bang?

In this chapter, we share our view that quantum mechanics is in fact deterministic, local and realistic, in complete contradiction with most people's perception of Bell's theorem, thanks to our new theory of *parallel lives*. Accordingly, what we perceive as the so-called "collapse of the wavefunction" is but an illusion. Then we ask the fundamental question: Can a purely deterministic Quantum Theory give rise to the *illusion* of nondeterminism, randomness, probabilities, and ultimately can free will emerge from such a theory?

Keywords Free will • Realism • Locality • Church of the larger Hilbert space • Determinism • Parallel lives

G. Brassard (✉) • P. Raymond-Robichaud
Département IRO, Université de Montréal, C.P. 6128, Succ. Centre–Ville,
Montréal (QC), H3C 3J7, Canada
e-mail: brassard@iro.umontreal.ca; paulrr16@hotmail.com

A. Suarez and P. Adams (eds.), *Is Science Compatible with Free Will?: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience*,
DOI 10.1007/978-1-4614-5212-6_4,

4.1 Introduction

By the end of the nineteenth century, most physicists had evolved a completely deterministic view of the world. Even though he had many precursors, such as Paul Henri Thiry, Baron d'Holbach [1770], with his very influential *Système de la Nature*, it was the great French mathematician and astronomer Pierre-Simon, marquis de Laplace [1814], who expressed in the clearest terms the philosophy according to which everything is predetermined by the initial conditions. In his *Essai philosophique sur les probabilités*, he wrote:

> We ought to regard the present state of the universe as the effect of its anterior state and as the cause of the one which is to follow. Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it—an intelligence sufficiently vast to submit these data to analysis—it would embrace in the same formula the movements of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes.

If Laplace were right, would there be any possibility for conscious beings to exercise free will? Anything we might imagine that we are deciding would in fact have been "written" from the initial conditions existing at the time of the Big Bang! It would seem that free will requires some form of nondeterminism or randomness; that it cannot take hold unless some events happen without a cause.[1] Even though chaos theory makes it impossible to predict the future in a fully deterministic Universe as soon as there is even the tiniest imprecision on the initial conditions, these initial conditions would exist precisely according to classical physics, and thus the future would be determined, independently of our possibility of predicting it.

In the twentieth century, quantum mechanics ushered in one of the greatest revolutions in the history of science. In particular, it is generally considered to be *the* ultimate theory capable of explaining the emergence of randomness by virtue of a mysterious process known as the "collapse of the wavefunction", which seems to be inherent to irreversible quantum *measurements*. Therefore, quantum mechanics can be thought of as God's wonderfully imaginative solution to the problem of providing His creatures with free will in an otherwise well-ordered Universe. Nevertheless, Einstein so disliked the idea of true randomness in Nature that he claimed to be "convinced that *He* [God] does not throw dice" in a 1926 letter to Born [Einstein et al. 1971]. Most physicists today would say that Einstein was wrong in rejecting the occurrence of truly random events. But was he?

In this chapter, we share our view that quantum mechanics is in fact deterministic, local and realistic, in complete contradiction with most people's perception of Bell's theorem, thanks to our new theory of *parallel lives*. Accordingly, what we

[1] Nevertheless, we do acknowledge that *compatibilists* hold the belief that free will and determinism are compatible ideas, and that it is possible to believe both without being logically inconsistent. See http://plato.stanford.edu/entries/compatibilism/, accessed on 29 February 2012.

perceive as the so-called "collapse of the wavefunction" is but an illusion. Then we ask the fundamental question: Can a purely deterministic quantum theory give rise to the *illusion* of nondeterminism, randomness, probabilities, and ultimately can free will emerge from such a theory?

For the sake of liveliness, the nontechnical style of this chapter is purposely that of a spontaneous after-dinner speech. It is meant for the enjoyment of a curiosity-driven and scientifically minded readership who does not have prior knowledge in quantum mechanics. Occasional remarks and more rigorous details for the benefit of the expert are offered in the footnotes with no apologies for the casual reader. The next three sections review the standard notions of pure and mixed states, of entanglement, and of how one part of an entangled state can be described. Readers familiar with these notions may prefer to proceed directly with Sect. 4.5, which describes the Church of the Larger Hilbert Space, a central notion to this chapter since it restores determinism into quantum mechanics. Section 4.6 attempts to go one step further in restoring also locality at the expense of realism, but it fails to do so. Then, Sect. 4.7 announces our new theory, which we call *parallel lives*, in which both locality and realism are restored in a physical world in which Bell's inequalities are nevertheless violated. Finally, Sect. 4.8 discusses the implication of all of the above on the existence or not of free will, be it real or illusory.

4.2 Pure and Mixed States

According to quantum mechanics, one has to distinguish between pure and mixed states. A *pure* state, generally denoted $|\Psi\rangle$ following Paul Dirac, is used to represent a state about which everything is known. For instance, $|0\rangle$ and $|1\rangle$ correspond to the classical notion of bits 0 and 1, whereas $|\Psi\rangle = \frac{1}{\sqrt{2}}|0\rangle + \frac{1}{\sqrt{2}}|1\rangle$ denotes a *qubit* (for "quantum bit"), whose state is an equal *superposition* of $|0\rangle$ and $|1\rangle$. This means that $|\Psi\rangle$ represents a state that corresponds *simultaneously* to classical bit values 0 and 1, each with *amplitude* $\frac{1}{\sqrt{2}}$. If this qubit is measured in the so-called *computational basis* ($|0\rangle$ vs. $|1\rangle$), standard quantum mechanics has it that it will *collapse* to either classical state $|0\rangle$ or $|1\rangle$, each with a probability given by the square of the norm of the corresponding amplitude, here $\left|\frac{1}{\sqrt{2}}\right|^2 = 1/2$ for each alternative. Even though the specific result of the measurement is not determined by the pure state, and two strictly identical particles in that same state could yield different results following the same measurement, the probabilities associated with such measurement outcomes are known exactly. Furthermore, this particular state would behave in a totally deterministic manner if subjected to a *different* measurement, known in this case as the Hadamard measurement (or measurement in the Hadamard basis "H"), which asks it to "choose" between $\mathsf{H}|0\rangle = \frac{1}{\sqrt{2}}|0\rangle + \frac{1}{\sqrt{2}}|1\rangle$ and $\mathsf{H}|1\rangle = \frac{1}{\sqrt{2}}|0\rangle - \frac{1}{\sqrt{2}}|1\rangle$. In this case, our state would choose the former since in

fact $|\Psi\rangle = \mathsf{H}|0\rangle$. Peres [1995] *defined* a pure state as one for which there exists a complete measurement (which he calls a "maximal test") under which it behaves deterministically.

In contrast, *mixed* states are used when there is intrinsic uncertainty not just about the result of some measurement but about the result of *all* possible complete measurements, hence about the state itself. One way to picture a mixed state is to think of a black box inhabited by a Daemon. When a user pushes a button, the Daemon spits out a state that it chooses at random,[2] say with equal probability between $|0\rangle$ and $|1\rangle$. Such a mixed state would be denoted as

$$\mathcal{E}_1 = \left\{\left(|0\rangle, {}^1\!/_2\right), \left(|1\rangle, {}^1\!/_2\right)\right\}. \tag{4.1}$$

More generally, a mixture of k different pure states is denoted as

$$\mathcal{E} = \{(|\Psi_1\rangle, p_1), (|\Psi_2\rangle, p_2), \ldots, (|\Psi_k\rangle, p_k)\} = \{(|\Psi_i\rangle, p_i)\}_{i=1}^{k}, \tag{4.2}$$

which means that the Daemon chooses some $|\Psi_i\rangle$ with probability $p_i, 1 \leq i \leq k$, where the probabilities sum up to 1. It is legitimate to wonder if such a state is pure since the Daemon knows which $|\Psi_i\rangle$ it chose, or if it is mixed since the user does not know. In a sense it is both. Nevertheless, no measurement chosen by the user will provide a deterministic answer. For instance, a measurement of $\mathcal{E}_1$ in the computational basis will reveal the Daemon's random choice, which has equal probability ${}^1\!/_2$ of being $|0\rangle$ or $|1\rangle$. On the other hand, a measurement in the Hadamard basis will produce $\mathsf{H}|0\rangle$ or $\mathsf{H}|1\rangle$ with equal probability ${}^1\!/_2$ since such would be the case regardless of whether the Daemon had spit out $|0\rangle$ or $|1\rangle$. Thus we see that the randomness lies with the Daemon in one case and with the user's measurement in the other case, but the final result is the same. More interestingly, it can be demonstrated that *any* measurement on $\mathcal{E}_1$ that would ask it to choose between two arbitrary one-qubit orthogonal states would choose either one with equal probability. (Two states are *orthogonal* if it is possible in principle to distinguish perfectly between them, such as $|0\rangle$ and $|1\rangle$, or $\mathsf{H}|0\rangle$ and $\mathsf{H}|1\rangle$). By Peres' definition, $\mathcal{E}_1$ is not pure since there does not exist a complete measurement under which it behaves deterministically.

Mixed states can be described as above by a mixture of pure states, but they can also be described in two other ways. One of them is known as the *density matrix* (aka density *operator*). This is a matrix (an array of numbers) that can be calculated mathematically from the more intuitive mixture $\{(|\Psi_i\rangle, p_i)\}_{i=1}^{k}$ of pure states. The remarkable fact about density matrices is that different mixtures can give rise to the same matrix, yet this matrix represents *all* that is measurable about the state, by any measurement whatsoever "allowed" by quantum mechanics. For instance, the

[2] This must be a *true* random choice, possibly implemented by a quantum-mechanical process; flipping a classical coin would not suffice here.

density matrix that is computed from (4.1) is identical to that arising from the apparently different mixture

$$\mathcal{E}_2 = \left\{ \left(\mathsf{H}|0\rangle, {}^1\!/_2 \right), \left(\mathsf{H}|1\rangle, {}^1\!/_2 \right) \right\}. \tag{4.3}$$

In other words, if we trust a Daemon to send us an equal mixture of $|0\rangle$ and $|1\rangle$ (4.1) but in fact it provides us with an equal mixture of $\mathsf{H}|0\rangle$ and $\mathsf{H}|1\rangle$ (4.3), we shall never be able to notice that it is "cheating"![3] Given that these two mixtures are impossible to distinguish, it makes sense to consider the corresponding mixed states as actually *identical*. Just for completeness, notice that even mixtures featuring more than two pure states can be indistinguishable from those above. For instance, mixture

$$\mathcal{E}_3 = \left\{ \left(|0\rangle, {}^1\!/_3 \right), \left(\frac{1}{2}|0\rangle + \frac{\sqrt{3}}{2}|1\rangle, {}^1\!/_3 \right), \left(\frac{1}{2}|0\rangle - \frac{\sqrt{3}}{2}|1\rangle, {}^1\!/_3 \right) \right\} \tag{4.4}$$

is indistinguishable from (hence identical to) mixtures $\mathcal{E}_1$ and $\mathcal{E}_2$ because it gives rise to the same density matrix.

We postpone until Sect. 4.5 discussion of the third way—by far the most interesting—in which one may think of mixed states.

4.3 Entanglement

The concept of *entanglement* was first published (although not named) by Einstein, in collaboration with Podolsky and Rosen, in a failed attempt to demonstrate the *incompleteness* of the quantum formalism [Einstein et al., 1935]. However, there is historical evidence that the notion had been anticipated by Schrödinger several years previously, who was quick to understand the importance of entanglement: "I would not call that *one* but rather *the* characteristic trait of quantum mechanics, the one that enforces its entire departure from classical lines of thought" [Schrödinger, 1935]. We could not agree more with this assessment. Our quantum world is not classical *because*, as spectacularly demonstrated by Bell [1964], entanglement can*not* be explained by any classical local realistic theory of the sort that was so dear to Einstein. (Or can it? We'll come back to this question in Sect. 4.7.) Indeed, we consider entanglement to be *the* key to understanding Nature. We would even go so far as to say that it's our best window into probing the soul of the Universe. The other nonclassical aspects of quantum mechanics, such as the

[3] A much more remarkable example of cheating is possible for a Daemon who would be "paid" to produce randomly chosen Bell states. It could produce instead pairs of purely classical uncorrelated random bits. These mixtures being identical in terms of density matrices, such cheating would be strictly undetectable by the user. This is profitable for the Daemon because classical bits are so much easier to produce than Bell states!

quantization of energy and its consequence on the photoelectric effect—which earned Einstein his Nobel Prize in 1921—are no doubt important, but lag far behind the magic of entanglement on our personal wonder scale.

Entanglement is a phenomenon by which two (or more) physically separated systems must sometimes be thought of as a single (nonlocal) entity. The simplest example of entanglement is known as the *singlet state*,

$$|\Psi^-\rangle = \frac{1}{\sqrt{2}}|01\rangle - \frac{1}{\sqrt{2}}|10\rangle, \tag{4.5}$$

which consists of two particles. A measurement of both particles in the computational basis results in either outcome $|01\rangle$ or $|10\rangle$, each with equal probability since $|\pm\frac{1}{\sqrt{2}}|^2 = {}^1/_2$. Here, outcome $|01\rangle$ means that the first particle is measured as $|0\rangle$ and the second as $|1\rangle$, and similarly for outcome $|10\rangle$. In other words, the two particles yield opposite answers when they are measured in the computational basis. So far, this is not more mysterious than if someone had flipped a penny, sliced it through its edge, put each half-penny in an envelope, and mailed the envelopes to two distant locations. When the envelopes are opened ("measured"), there is no surprise in the fact that each one reveals a seemingly random result (heads or tails) but that the two results are complementary. Such an explanation would correspond to purely classical *mixed* state

$$\mathcal{E}_0 = \{(|01\rangle, {}^1/_2), (|10\rangle, {}^1/_2)\}, \tag{4.6}$$

where $|0\rangle$ stands for heads and $|1\rangle$ stands for tails.

What makes this singlet state so marvellous is that quantum mechanics asserts that $|\Psi^-\rangle$ is indeed the pure state given in (4.5) and not the mixed state of (4.6), and that those are very different indeed. In particular, the result of *any* measurement is *not* predetermined (as it would be with the half-penny analogy): it comes into existence only as a result of the measurement itself. This is particularly mysterious when the two particles are arbitrarily far apart because it is as if they were magic coins which, when flipped, always provide opposite, yet *freshly* random, outcomes. In fact, the two particles provide opposite answers to *any* complete measurement, provided they are subjected to the same one. It's like an old couple who disagrees on any question you may ask them... even when they don't have a clue about the answer and hence respond randomly! This phenomenon can be "explained" by elementary linear algebra, according to which state $|\Psi^-\rangle$, as given in (4.5), is *mathematically* equivalent to

$$|\Psi^-\rangle = \frac{1}{\sqrt{2}}|\psi\rangle|\phi\rangle - \frac{1}{\sqrt{2}}|\phi\rangle|\psi\rangle \tag{4.7}$$

for *any* two one-qubit orthogonal states $|\psi\rangle$ and $|\phi\rangle$, such as $\mathsf{H}|0\rangle$ and $\mathsf{H}|1\rangle$. It is important to understand that this behaviour would *not* occur with the mixed state of

(4.6) because, in that case, asking the two particles to "choose" between $\mathsf{H}|0\rangle$ and $\mathsf{H}|1\rangle$ would produce two random and *uncorrelated* outcomes.

An entangled state such as the singlet behaves exactly *as if* the first particle, when asked by a measurement to choose between orthogonal states $|\psi\rangle$ and $|\phi\rangle$, flipped a fair coin to decide which one to select, and then "instructed" the other particle to *instantaneously* assume the opposite state. This gives the *impression* of instantaneous action at a distance, a concept that so revolted Einstein that he derisively called it *spukhafte Fernwirkungen* ("Spooky action at a distance"). But is this really what happens or is it only a naïve "explanation"? We shall come back to this most fundamental issue in Sects. 4.6 and 4.7.

It has been experimentally demonstrated that if indeed the particles had to communicate, then the effect of the first measurement on the second particle would have to take place at least 10,000 times faster than at the speed of light [Salart et al., 2008]. Even more amazingly, *relativistic* experiments have been performed, following a fascinating theoretical proposal by Suarez and Scarani [1997], in which the predictions of quantum mechanics continue to hold even when the two particles move apart quickly enough that they are both measured before the other in their respective inertial reference frames [Stefanov et al., 2002]. These remarkable experiments make it untenable to claim that the first measured particle somehow sends a signal to tell the other how to behave. This has prompted Gisin [2013] to assert that quantum correlations "emerge from outside space–time".

We highly recommend the exceptionally lucid and entertaining popular accounts of some classically impossible marvels made possible by entanglement that have been written by Mermin [1981, 1994] for the *American Journal of Physics*.

4.4 Describing One Part of an Entangled State

The defining characteristic of a pure entangled state split between two distant locations is that neither of the local subsystems can be described as a pure state of its own. This should be clear from Peres' definition of a pure state and the fact that each part of an entangled state is so that its outcome is not predetermined, no matter to which complete measurement it is subjected. Nevertheless, it makes sense to wonder if there is a way to describe the state of one of the subsystems.

One natural approach is to see what would happen if we measured the *other* subsystem. Consider for instance the singlet state $|\Psi^-\rangle$ and let us measure one of the particles in the computational basis. We have seen that the outcome is $|0\rangle$ (resp. $|1\rangle$) with probability $^1/_2$, in which case the other system is now in state $|1\rangle$ (resp. $|0\rangle$). Therefore, if one system is measured *and one forgets the outcome of the measurement*, the unmeasured system is in state $|0\rangle$ with probability $^1/_2$ and in state $|1\rangle$ also with probability $^1/_2$. In other words, this system is in mixed state $\mathcal{E}_1$, according to (4.1). But the first system could have been measured in the Hadamard basis instead. Depending on the result of this measurement, the unmeasured system would then be left either in state $\mathsf{H}|0\rangle$ or $\mathsf{H}|1\rangle$, each with probability $^1/_2$. If we forget again the result

of the measurement, the unmeasured system is therefore in mixed state $\mathcal{E}_2$, according to (4.3). Now, remember that mixtures $\mathcal{E}_1$ and $\mathcal{E}_2$ are considered to be identical since they give rise to the same density matrix. More generally, it can be demonstrated from the formalism of quantum mechanics that no matter which complete measurement is performed on one subsystem of an arbitrary entangled state, the other subsystem always ends up in the same mixed state in terms of a density matrix, albeit not necessarily according to the same mixture of pure states.

It follows that nothing can be more natural than to describe one subsystem of an entangled state by the mixed state in which this subsystem *would* be left *if* the *other* subsystem were measured. This is well-defined since the resulting density matrix does not depend on how the other subsystem is measured. If we carry this reasoning to its inescapable conclusion, it makes sense to describe the state of a subsystem in this way *even if the other subsystem has not been measured yet*, indeed even if it is *never* to be measured. When we consider the state of a subsystem of an entangled state in this way, we say that we *trace out* the other subsystem.

Section 4.3 may have left you with the impression that entanglement requires instantaneous communication, which would be incompatible with Einstein's special theory of relativity. If we remember that the density matrix describes all that can be measured about a quantum system, however, it follows from the above discussion that entanglement *cannot* be used to *signal* information between two points in space since no operation performed on one subsystem of an entangled state can have a *measurable* effect on the other subsystem. It is as if quantum systems were capable of instantaneous communication, but only in tantalizing ways that could not be harnessed by us, macroscopic humans, to establish such communication between ourselves. We shall come back on the consequences of this crucial issue in Sects. 4.6 and 4.7.

4.5 Church of the Larger Hilbert Space

We have just seen that the state of any subsystem of a pure (or, for that matter, mixed as well) entangled state can be expressed as a mixed state in a unique and natural way. It is remarkable that the converse holds. We saw in Sect. 4.2 that mixed states can be described either as mixtures of pure states (possibly under the control of a Daemon) or as density matrices, but we promised a third way and here it is. *Any* mixed state can be described as the trace-out of some subsystem from an appropriate *pure* state. Such a pure state is called a *purification* of the mixed state under consideration.

There is an easy way (theoretically speaking) to construct a purification of an arbitrary mixture $\mathcal{E} = \{(|\Psi_i\rangle, p_i)\}_{i=1}^{k}$. For this, consider some other quantum system that could be in any of k orthogonal states $|\Phi_1\rangle, |\Phi_2\rangle, \ldots, |\Phi_k\rangle$ and consider pure state

$$|\Psi\rangle = \sum_{i=1}^{k} \sqrt{p_i}\,|\Psi_i\rangle\,|\,\Phi_i\rangle,$$

where "$\sum_{i=1}^{k}$" serves to denote a quantum superposition on k pure states. If the right-hand subsystem of $|\Psi\rangle$ were measured by asking it to "choose" between one of the $|\Phi_i\rangle$'s, each $|\Phi_i\rangle$ would be chosen with probability $|\sqrt{p_i}|^2 = p_i$, leaving the unmeasured left-hand subsystem in state $|\Psi_i\rangle$.

Now, imagine that it were our friend the Daemon who prepared pure state $|\Psi\rangle$ and measured its right-hand subsystem. By learning which $|\Phi_i\rangle$ is obtained, with probability p_i, the Daemon would know in which *pure* state $|\Psi_i\rangle$ the unmeasured subsystem is. If the Daemon spits out this subsystem to the user, without revealing the result of the measurement, the user receives a *mixed* state corresponding to mixture $\mathcal{E}$. As in Sect. 4.2, this system is in a pure state for the Daemon and in a mixed state for the user. The beauty of this concept is that it works even if the Daemon has not, in fact, measured the right-hand subsystem of the pure state it had created. Even better, it still works if the Daemon has destroyed that right-hand subsystem, inasmuch as a quantum state can be destroyed, to prevent any temptation to measure it later and sell the answer to the user! In this case, the surviving quantum system would be in mixed state $\mathcal{E}$ not only for the user but also for the Daemon.

The fact that any mixed state can be considered as the trace-out of one of its purifications is the fundamental tenet of the Church of the Larger Hilbert Space, a term coined by John Smolin because the formalism of quantum mechanics has pure quantum states inhabit so-called Hilbert spaces and any mixed state can be thought of as a subsystem from a pure state than lives in a *larger* Hilbert space.

Everything that we have explained so far in this chapter corresponds to strictly orthodox quantum mechanics and no (serious) physicist would disagree with a single word from it. From this point on, however, we articulate our personal beliefs concerning the world in which we live, which are admittedly very similar to the "relative state" formulation of quantum mechanics put forward by Everett [1957] more than 50 years ago; see also Byrne [2007].

The *weak* Faithfuls in the Church of the Larger Hilbert Space believe in the fact that any mixed state can be *thought of* as the trace-out of some imaginary purification, but this is only for mathematical convenience. In fact, it is not possible to believe in the predictions and formalism of quantum mechanics without being (at least) a weak faithful since the (mathematical) existence of a purification for any mixed state is a *theorem* that can be derived from first principles.

The *strong* Faithfuls—among whom we stand—believe that to any mixed state that actually exists, there corresponds somewhere in the Universe an appropriate purification. This is an extremely far-reaching belief since it implies (among other things) that the "collapse of the wavefunction", which orthodox quantum mechanics associates with measurements, is but an illusion. In fact, strong belief in the Church implies that quantum mechanics is strictly unitary and therefore reversible. If we forget for simplicity the necessity to apply relativistic corrections, the Universe is

ruled by one law only, known as Schrödinger's equation. This equation is deterministic—even linear—and therefore so is the entire evolution of the Universe.

Let us consider for instance the simplest case of orthodox collapse of the wavefunction: the measurement of a single diagonally polarized photon (a particle of light) by an apparatus that distinguishes between horizontal and vertical polarizations. For definiteness, consider a calcite crystal that splits an incoming light beam between horizontally and vertically polarized sub-beams followed by two single-photon detectors (which we assume perfect for sake of the argument). Any horizontally polarized photon would cause one of the two detectors to react, whereas a vertically polarized photon would cause the other detector to react. According to orthodox quantum mechanics, a diagonally polarized photon would hit the crystal and then continue in quantum superposition of both paths until it hits both detectors. At this point, one (and only one) of the detectors would "see" the photon and produce a macroscopic effect that would be detectable by the (human) observer. For some, the phenomenon would become irreversible as soon as it has had a macroscopic effect inside the detector; for others only when some observer becomes conscious of the outcome.

According to the Strong Church of the Larger Hilbert Space, neither is the case: the diagonally polarized photon is in fact in an equal superposition of being horizontally and vertically polarized (so far, this is in strict accordance with orthodox quantum mechanics) and the crystal merely puts the photon in a superposition of both the horizontally and vertically polarized paths (still in accordance with orthodox quantum mechanics). But when the photon hits both detectors, it becomes *entangled* with them. The composite system photon-detectors is now in an equal superposition of the photon being horizontally polarized and the horizontal-polarization detector having reacted with the photon being vertically polarized and the vertical-polarization detector having reacted. And when the observer looks at the detectors, he or she becomes entangled with the photon-detector system so that now the photon-detector-observer system is in an equal superposition of the photon being horizontally polarized, the horizontal-polarization detector having reacted and the observer having seen the horizontal-polarization detector reacting with the same events corresponding to a vertically polarized photon.

From this perspective, there is no collapse. The horizontal detection is as real as the vertical one. But any (human) observer becomes aware of only one outcome, and here lies the *apparent* paradox. If indeed both events occur (in quantum superposition), how come our experience makes us (humans) believe that only one outcome (apparently chosen at random by Nature) has actually occurred? In his groundbreaking paper, Everett [1957] proposed the following analogy:

> Arguments that the world picture presented by this theory is contradicted by experience, because we are unaware of any branching process, are like the criticism of the Copernican theory that the mobility of the earth as a real physical fact is incompatible with the common sense interpretation of nature because we feel no such motion. In both cases the argument fails when it is shown that the theory itself predicts that our experience will be what it in fact is. (In the Copernican case the addition of Newtonian physics was required to be able to show that the earth's inhabitants would be unaware of any motion of the earth.)

In other words, it is not because we (humans) cannot feel the Earth moving under our feet that it stands still at the centre of the Universe! Similarly, it is not because we cannot feel the universal superposition that it does not exist. According to the Strong Church of the Larger Hilbert Space, the Earth as we feel it has but a tiny amplitude in the Universal wavefunction, and each one of us has an even tinier amplitude. This perspective is very humbling indeed, much more so than accepting the insignificance of the Earth within the classical Universe, but this is nevertheless the perspective in which we most passionately believe.

It remains to see how the Church of the Larger Hilbert Space can account for the phenomenon described in Sect. 4.3 when we discussed the measurement of far-apart entangled particles.[4] Consider again two particles in the singlet state (4.5) and assume that they are both subjected to the same measurement, which asks them to "choose" between orthogonal states $|\psi\rangle$ and $|\phi\rangle$. We can think of the two particles as being in the state given by (4.7), which once again is mathematically equivalent to (4.5), and initially the measurement apparatuses have not reacted, so that they are not entangled with the particles. The joint state of the apparatus–particle–particle–apparatus system can therefore be described as

$$|?\rangle\left(\frac{1}{\sqrt{2}}|\psi\rangle|\phi\rangle - \frac{1}{\sqrt{2}}|\phi\rangle|\psi\rangle\right)|?\rangle, \tag{4.8}$$

where $|?\rangle$ represents a measurement apparatus that has not yet reacted. This is mathematically equivalent to

$$\frac{1}{\sqrt{2}}|?\rangle|\psi\rangle|\phi\rangle|?\rangle - \frac{1}{\sqrt{2}}|?\rangle|\phi\rangle|\psi\rangle|?\rangle. \tag{4.9}$$

Let us say without loss of generality that the particle on the left is measured first. According to the Church of the Larger Hilbert Space, this has the effect of entangling it with its measurement apparatus. However (and contrary to the teachings of standard quantum mechanics), the two particles remain entangled, albeit no longer in the singlet state. Now, the joint state of the complete system has (unitarily) evolved to

$$\left(\frac{1}{\sqrt{2}}|\Psi\rangle|\psi\rangle|\phi\rangle - \frac{1}{\sqrt{2}}|\Phi\rangle|\phi\rangle|\psi\rangle\right)|?\rangle, \tag{4.10}$$

[4] Of course, we must account for all the nonclassical correlations that violate various forms of Bell inequalities, not only for the (classically explicable) fact that two particles in the singlet state will always give opposite answers when subjected to the same measurement. This paragraph can be adapted *mutatis mutandis* to any pair of measurements, including POVMs, on an arbitrary bipartite entangled state, as well as to similar scenarios for multipartite entanglement.

where $|\Psi\rangle$ (resp. $|\Phi\rangle$) represents the state of a measurement apparatus that has registered a particle in state $|\psi\rangle$ (resp. $|\phi\rangle$). Note that the apparatus on the right is still unentangled with the other systems under consideration. Finally, when the particle on the right is measured, the system evolves to

$$\frac{1}{\sqrt{2}}|\Psi\rangle|\psi\rangle|\phi\rangle|\Phi\rangle - \frac{1}{\sqrt{2}}|\Phi\rangle|\phi\rangle|\psi\rangle|\Psi\rangle\,. \tag{4.11}$$

At this point, if we trace out the two particles, the detectors are left in *mixed* state

$$\left\{\left(|\Psi\rangle|\Phi\rangle, {}^1\!/_2\right), \left(|\Phi\rangle|\Psi\rangle, {}^1\!/_2\right)\right\}, \tag{4.12}$$

which is exactly as it should: they have produced random but complementary outcomes. Naturally, we could also involve two human observers in this scenario. If we had, they would enter the macroscopic entangled state of (4.11); at that point, they would in a superposition of having seen the two possible complementary sets of outcomes, but they would be blissfully unaware of this.

As an amusing anecdote, we cannot resist mentioning the (real-life!) venture called cheap universes.[5] For a mere $3.95, or unlimited use for $1.99 on an iPhone, you can select two courses of action (such as "I shall either go on a hike, or I shall take a nice hot bath") and ask cheap universes to make a purely quantum choice between the two alternatives.[6] Provided you have self-pledged to obey the outcome, you may proceed lightheartedly because you know that you are also performing the other action in the Universal wavefunction. Indeed, should the consequences of having indulged in a nice hot bath turn out to be disastrous, you can take comfort in knowing that you have *also* gone on a hike and hope that this was indeed the path to happiness. Sounds crazy? Not to us!

The Strong Church of the Larger Hilbert Space is different from (but not incompatible with) the so-called Many-World Interpretation of Quantum Mechanics (usually associated with the name of Everett) in the sense that we believe in a single Universe—not in the "Multiverse" advocated by the Many-World Interpretation followers—but one in which quantum mechanics rules at face value: We (poor humans) perceive only what we call the classical states, but arbitrarily complex superpositions of them do in fact exist in reality.

[5] http://www.cheapuniverses.com, accessed on 29 February 2012.

[6] Specifically, cheap universes uses a commercial device called QUANTIS, available from ID Quantique, in which "photons are sent one by one onto a semi-transparent mirror and detected; the exclusive events (reflection/transmission) are associated to '0'/'1' bit values". See http://www.idquantique.com/true-random-number-generator/products-overview.html, accessed on 29 February 2012. According to our example, we would associate outcome 0 with "I shall go on a hike" and outcome 1 with "I shall take a nice hot bath".

4.6 Can Locality be Restored Outside of the Church?

Einstein thought that quantum mechanics must be *incomplete* because it did not fulfil his wish for a *local* and *realistic* theory. We distinguish between *strong* realism, according to which any property of a physical system registered by a measurement apparatus (or by any other process by which the system is observed) existed prior to the measurement, so that the apparatus merely reveals what was already there, and *weak* realism, according to which a physical system can respond probabilistically to a measurement apparatus, but the probability distribution of the possible outcomes exists prior to the measurement. For instance, the diagonal polarization of a horizontally polarized photon exhibits weak realism according to quantum mechanics because its measurement behaves randomly, yet with well-defined probabilities (in this case with equal probability of registering a $+45^\circ$ or a -45° outcome).

Similarly, we distinguish between *strong* locality, according to which no action performed at point A can have an effect on point B faster than the time it takes light to go from A to B, and *weak* locality, according to which there can be no *observable* such effect. As we have seen already, if two particles are jointly in the singlet state (4.5) and if one is measured, yielding outcome $|0\rangle$, the other particle behaves as if its state had instantaneously changed from being half a singlet to pure state $|1\rangle$, no matter how far apart the two particles are. Even though this phenomenon *seems* to violate strong locality (we shall come back on this issue below and in the next section), it is important to understand that it does *not* violate weak locality because the instantaneous effect (if it exists) cannot be detected by any process allowed by quantum mechanics. Taking account of the special theory of relativity, violations of weak locality would enable reversals in causality (effects could precede causes), whereas violations of strong locality have no such spectacular consequences. Fortunately, quantum mechanics does not allow *any* violation of weak locality. From now on, "locality" will be understood to mean "strong locality" unless specified otherwise.

For a strong faithful in the Church of the Larger Hilbert Space, the issue of locality can take different flavours. At one extreme, the wavefunction is the one and only reality and the question does not even make sense. The Universe *is* in a massive superposition and anything that appears to involve a random quantum choice in one branch of the superposition "simply" makes the universal superposition more complicated; the issue of locality does not even spring up. This position is often considered to be a "cop out" by those who are not faithfuls of the Church, who think that believers are simply avoiding the issue rather than trying to explain it. The other extreme among the faithfuls is populated by the advocates of the many-world interpretation of quantum mechanics, some of whom consider that the entire world splits up each time a random quantum choice appears to be made. Such a split is highly nonlocal if it is instantaneous. In the next section, we shall present our *parallel lives* interpretation of the Church of the Larger Hilbert Space, which is

fully compatible with locality. In the rest of this section, however, we shall step outside of the Church and attempt to reconcile locality with quantum mechanics while denying the possibility for macroscopic objects (such as human observers) to enter into a superposition.

The great discovery of Bell [1964], or so it seems, is that the predictions of quantum mechanics are incompatible with any possible strongly local and weakly realistic theory of the world.[7] Since quantum mechanics has been vindicated by increasingly sophisticated experiments [Freedman and Clauser, 1972, Aspect et al., 1981, 1982a, 1982b, Stefanov et al., 2002, Salart et al., 2008, etc], most physicists infer that there is no other choice but to forego locality. However, we beg to disagree on the inevitability of this conclusion. If the world cannot be simultaneously local and realistic, could locality be restored at the expense of realism?

One may attempt to achieve this by accepting that the state of a quantum system is fundamentally *subjective* (or to be technically more exact, *epistemic*). For instance, the same particle can be in one state for one observer and in a different state for another. *And both observers can be perfectly correct about the state of the particle!* However, they must have *compatible* beliefs in the sense that there must exist at least one pure state that is excluded by neither observer.[8]

For sake of the argument, consider again a quantum system in the singlet state (4.5) so that the two particles are arbitrarily far apart, say at points A and B, which are inhabited by Alice and Bob, respectively. We have seen that the state of either particle can be described locally by mixture $\mathcal{E}_1$ from (4.1). To stress that we are not talking about the *specific* mixture of pure states explicit in $\mathcal{E}_1$ (since the state of these particles can just as well be described by mixtures $\mathcal{E}_2$ or $\mathcal{E}_3$), let us denote the corresponding density matrix by ρ, which is uniquely defined.

Consider what happens if Alice measures her particle in the computational basis and obtains (say) outcome $|0\rangle$. Then, assuming Bob has not interacted with his particle, Alice *knows* that Bob's particle is no longer in mixed state ρ: now it is in state $|1\rangle$. But for Bob, nothing has changed! His particle was in state ρ before Alice's measurement and it *still* is in this same state immediately after the far-away measurement. In other words, the particle at point B is in state $|1\rangle$ for Alice and in state ρ for Bob, *and both observers are correct* in their assertions concerning the *same* particle. This is reminiscent of the proverbial Indian story of the blind men and an elephant.[9]

The effect of Alice's measurement *can* propagate to Bob, but *only* if a *classical* message transits between them. However, such a message cannot travel faster than at the speed of light. It follows that there is no faster-than-light change in the state of the particle at point B, as seen by Bob from that point. More generally, no operation performed at any point in space can have an instantaneous *observable* effect on any

[7] To be historically more accurate, Bell's original 1964 paper was concerned with strong realism only, but it can be strengthened to take account of weak realism.

[8] To be technically exact and much more general, there must exist at least one *ontic* state compatible with both *epistemic* beliefs, unless we are ready to accept that there is no underlying reality at all [Pusey et al., 2012].

[9] http://en.wikipedia.org/wiki/Blind_men_and_an_elephant, accessed on 29 February 2012.

other point. Seen this way, no cause can have an effect faster than at the speed of light, causality is not violated, and Einstein can rest in peace.

Naturally, it *is* possible for an observer to be wrong about the state of a particle. For instance, if Alice prepares a particle in state $|0\rangle$ and sends it to Bob, who is far away, and if Bob subjects it to a Hadamard transformation without telling Alice, then Alice may think that the particle is still in state $|0\rangle$ and be wrong since it is now in state $\mathsf{H}|0\rangle$. However, this is not in contradiction with the above: It is not because the same particle can be in two different states according to two different observers and that both can be correct that anybody who has some opinion about the state of a quantum system is necessarily right! For Alice to know the state of a far-away particle, even subjectively, she must know what has happened to it after it left her hands. We shall therefore consider for simplicity a bipartite scenario in which each party knows what the other party is doing.

Can we completely restore locality at the expense of realism with this line of approach? Unfortunately, there is a serious problem. Consider again the case of Alice and Bob sharing a singlet state and of Alice measuring her particle. We argued above that Alice and Bob can both be correct if Alice thinks of Bob's particle as being in state $|1\rangle$ whereas Bob thinks of it as being in mixed state ρ. But now, if Bob decides to measure his particle, and if indeed his belief that it is in state ρ were correct, there would be no *local* reason that would prevent him from registering outcome $|0\rangle$, which is indeed possible when measuring ρ since it can be thought of as mixed state $\mathcal{E}_1$ from (4.1). This would no longer be compatible with Alice's belief that Bob's particle is in state $|1\rangle$, even though each party knows what the other is doing. Furthermore, if they meet in the future and compare notes, Alice and Bob will register correlations that are not in accordance with quantum mechanics. (Please remember that this section is written under the assumption that neither Alice nor Bob can be in a superposition of having seen both results).

Does it follow that quantum mechanics cannot be explained by a local theory even if we are willing to forego realism? It turns out that we can have our cake and eat it too, provided we reintegrate the Church of the Larger Hilbert Space. Contrary to the prevalent belief, the laws of nature can be simultaneously local and realistic, and yet obey all the predictions of quantum mechanics. In order to reconcile this claim with Bell's impossibility proof, please consider the quotation of Bell's at the opening of this chapter and read on.

4.7 Parallel Lives

We shall fully develop our *parallel lives* theory for quantum mechanics in a subsequent paper. Here, for simplicity, we explain how local realism can be consistent with bipartite correlations that are usually considered to be even more nonlocal than those allowed by quantum mechanics. Specifically, we consider the so-called *nonlocal box* introduced by Popescu and Rohrlich [1994], which we illustrate with a tale that takes place in an imaginary world, i.e., in a toy model of an alternative Universe. Our Universe follows Einstein's special theory of relativity so

that it is possible to assert, according to the principle of weak locality, that some events cannot influence the outcome of other far-away events that are sufficiently simultaneous.

Imagine two inhabitants of this Universe, Alice and Bob, who travel very far apart in their spaceships. Each one of them is carrying a box that features two buttons, labelled 0 and 1, and two lights, one green and one red. Once they are sufficiently distant, Alice and Bob independently flip fair coins to decide which button to push on their boxes, which causes one light to flash on each box. The experiment is performed with sufficient simultaneity that Alice's box cannot know the result of Bob's coin flip (hence the input to Bob's box) before it has to flash its own light, even if a signal travelling at the speed of light left Bob's spaceship at the flip of his coin toss to inform Alice's box of the outcome, and vice versa.[10]

After several instances of this experiment, Alice and Bob meet again to compare their results. They discover to their amazement that they saw different colours when and only when they had both pushed the "1" button. In a local classical world that denies the Church of the Larger Hilbert Space, in which Alice and Bob cannot enter into a superposition (as in the previous section), it is easy to see that such boxes cannot exist. More precisely, the best box that can be built cannot produce such results with a probability better than 75%. In a quantum-mechanical world, we can do better by the magic of entanglement, but the success probability cannot exceed $\cos^2 \pi/8 \approx 85\%$ [Cirel'son, 1980], hence our imaginary world is not ruled by quantum mechanics either. This is fine: remember that the purpose of this scenario is not to suggest a model of our world, but rather to show that it is possible in a local realistic world to violate a Bell inequality.

What is the trick? Imagine that each spaceship lives inside a bubble. When Alice pushes one button on her box, her bubble splits into two parallel bubbles. Each bubble contains a copy of the spaceship and its inhabitant. Inside one bubble, Alice has seen the red light flash on her box; inside the other bubble, she has seen the green light flash. From now on, the two bubbles are living parallel lives. They cannot interact between themselves in any way and will never meet again. The same phenomenon takes place when Bob pushes one button on his box. Please note that Alice's action has strictly no instantaneous influence on Bob's bubble (or bubbles if he has already manipulated his box): this splitting into parallel lives is a strictly local phenomenon.

Let us consider what happens if Alice and Bob, each of whom now lives inside two parallel bubbles although they cannot feel it in any way,[11] decide to travel toward each other and meet again. (A similar scenario can be involved if they

[10] We are implicitly ruling out the local realistic theory of *superdeterminism* here, according to which there is no way to prevent the boxes from knowing which button is pressed on the other box, not because a signal travels quickly enough between the boxes, but because everything being deterministic, each box knows everything about the future, including which buttons will be pushed anywhere in the Universe. See http://en.wikipedia.org/wiki/Superdeterminism, accessed on 29 February 2012.

[11] Remember Everett's analogy with medieval criticism of the Copernican theory concerning the fact that we cannot feel the Earth move under our feet.

decide to use classical communication in order to compare notes, rather than travelling.) This is where magic[12] takes place: Each of the two bubbles that contains Alice is allowed to interact and see only a single bubble that contains Bob, namely the bubble that satisfies the conditions described above. Note that such a perfect matching is always possible. Furthermore, each bubble can "know" with which other bubble to interact provided it keeps a (local) memory of which button was pressed and which light flashed. In this way, each copy of Alice and Bob will be under the illusion of correlations that "emerge from outside space–time" [Gisin, 2013]. Yet these correlations take place fully within space and time, in a completely local realistic Universe. In our imaginary world, the Einstein-Podolsky-Rosen argument [Einstein et al., 1935] fails because whenever Alice pushes a button and can predict something about Bob, she is really predicting, not what is happening instantaneously at Bob's place, but how their various lives will meet and interact in the future.

Let us stress again that we are not claiming that our Universe actually works as described above, because it does not, according to Cirel'son [1980]. Our point is that it is generally recognized that nonlocal boxes of the sort we have described cannot exist in any local realistic world, *and this is false* according to our toy model. To be more dramatic, consider Bell's theorem, or more precisely its best-known incarnation, the CHSH inequality due to Clauser et al. [1969]. This inequality states that "in any classical theory [...] a particular combination of correlations[13] lies between -2 and 2" [Popescu and Rohrlich, 1994]. The original purpose of this inequality is that it is violated by quantum mechanics since the same "particular combination of correlations" is predicted to be equal to $2\sqrt{2}$, hence quantum mechanics cannot be explained by a "classical theory" of the sort considered by Clauser et al. [1969] to derive their inequality. As demonstrated by Popescu and Rohrlich [1994], this combination can as large as four without violating weak locality, and indeed it is equal to four in our toy model of the world.

Have we uncovered a fundamental mistake in the paper of Clauser et al. [1969]? Not at all! Bell's inequalities (including CHSH and those from Bell's original 1964 paper) are proved, indeed correctly, under the assumption that the classical world is a theory of *local hidden variables*. The confusion comes from the fact that this has been widely misinterpreted to mean that quantum mechanics rules out any local realistic explanation of the world. For instance, Nielsen and Chuang [2000] wrote in their book: "These two assumptions together are known as the assumptions of local realism. [...] The Bell inequalities show that at least one of these assumptions is not correct. [...] Bell's inequalities together with substantial experimental evidence now points to the conclusion that either or both of locality and realism must be dropped from our view of the world." Note that Nielsen and Chuang consider here both locality and realism to be of the strong type, but our parallel lives mechanism is purely deterministic, hence it is strongly realistic as well.

[12] Remember Arthur C. Clarke's Third Law: "Any sufficiently advanced technology is indistinguishable from magic"!

[13] Specifically, $E(A,B)+E(A,B')+E(A',B)-E(A',B')$; for detail, please see Eqs. (1) and (2) from Popescu and Rohrlich [1994].

The virtue of our toy model is to demonstrate in an exceedingly simple way that local realistic worlds can produce correlations that are demonstrably impossible in any classical theory based on local hidden variables. Therefore, it illustrates the importance of understanding the true meaning of Bell's theorem. Nevertheless, it begs the question: what about quantum mechanics? Can *it* be explained in a local realistic parallel lives scenario?

It turns out that the idea of quantum mechanics being local and realistic in a theory analogous to parallel lives was discovered in the twentieth century: it can be traced back at least to Deutsch and Hayden [2000]. Similar ideas were introduced subsequently by Rubin [2001] and Blaylock [2010]. The article of Deutsch and Hayden focused on locality without precisely formulating definitions of realism or what we have called parallel lives, but their mathematical structure was quite similar to what we propose here. Of course, a complete reformulation of quantum mechanics along these lines is significantly more technical and complicated than what is needed to "explain" nonlocal boxes, but the conclusion is that the Church of the Larger Hilbert Space can be interpreted to provide a fully deterministic, strongly local and strongly realistic interpretation of quantum mechanics, Bell's theorem notwithstanding. Indeed, this interpretation is not about parallel branches or parallel Universes in a multiverse, but rather about parallel lives, which is a purely local phenomenon.

We are currently working on a follow-up article that will provide much more detail about our parallel lives theory.

4.8 Free Will?

At this point, the fundamental question is "Can a purely deterministic quantum theory give rise to at least the *illusion* of nondeterminism, randomness, probabilities, and ultimately can free will emerge from such a theory"? Please note that this section is written at the first person as it reflects solely the opinion of the first author. The second author resolutely does not believe in free will and therefore his position is that neither determinism nor randomness would be able to enable it.

I cannot answer in a definitive way the question asked at the beginning of this section. Certainly, I acknowledge the difficulty of deriving the emergence of probabilities as mathematically inevitable from a quantum Universe in which all events occur unitarily according to the Church of the Larger Hilbert Space [Kent, 2010]. However, we are faced with *exactly* the same difficulty if the collapse of the wave function does occur, or even in a purely classical world [Duhamel and Raymond-Robichaud, 2011]. I also acknowledge the difficulty of deriving free will from probabilities, randomness and nondeterminism. Nevertheless, I am inhabited by an unshakable belief that free will, *if* it exists, cannot have another origin, with apologies to the compatibilists.

In his own chapter in this book, Gisin [2013] expresses his view that the Many-World Interpretation of Quantum Mechanics "leaves no space for free

will". I suspect that he would have the same opinion concerning the Strong Church of the Larger Hilbert Space. [He also maintains that free will is not incompatible with the deterministic physics of Newton, but I fail to understand how classical physics could escape the "*intelligence*" of Laplace [1814], for which "nothing would be uncertain and the future, as the past, would be present to its eyes".] In any case, I admit that Gisin may be right, but my most fundamental disagreement lies deeper than physics or mathematics when he says: "I enjoy free will much more than I know anything about physics". I respect his opinion, but my personal position is that I would prefer to live in a world without free will rather than in one in which the wavefunction collapses nonunitarily. After all, lack of free will in a deterministic Universe does not deprive us from our capacity to experience surprise and find wonder in the world, because we cannot calculate, and hence predict, the future. But of course, whether or not free will exists, it does not extend to the point of letting each one of us choose in which of these two Universes we actually live!

Perhaps cheap universes is our ultimate window on free will. Provided we firmly decide to follow whichever course of action it chooses for us, we are free to populate both branches of the Universal superposition. In whichever branch we perceive ourselves to be, we have made the free choice of letting quantum phenomena decide for us. Of course, I am not seriously suggesting that free will did not exist until the inception of cheap universes, just as Bell [1990] was not serious when he asked if "the wavefunction of the world [was] waiting to jump for thousands of millions of years until a single-celled living creature appeared? Or did it have to wait a little longer, for some better qualified system. . . with a PhD?"!

4.9 Conclusion

In this essay, we have penned down for the first time our beliefs concerning the Universe in which we live, even though one of us (Brassard) has been inhabited by these thoughts for several decades. The more time goes by, the more convinced we are that they constitute the most rational explanation for our quantum world. We reject violently the notion that there would be a quantum–classical boundary and that physics is discontinuous, with a reversible (even unitary) evolution at the microscopic level but an irreversible collapse at the macroscopic level of measurements. It may be that free will can at best be an illusion in a world ruled by the Strong Church of the Larger Hilbert Space because every time you think that you make a decision (provided you use the services of cheap universes or some other source of true quantum randomness to make your choices), you also make the complementary decision in the universal superposition. However, what does it matter if free will does not truly exist, provided the illusion is perfect?[14]

[14] Seriously, we would not want to live in the Matrix imagined by the Wachowski siblings, no matter how perfect is the simulation. So, perhaps we *do* care after all!

We give the last words of wisdom to Bell [1990], who ended his Summary of "Against 'measurement' " by:

> I mean [. . .] by serious, that apparatus should not be separated off from the rest of the world into black boxes, as if it were not made of atoms and not ruled by quantum mechanics.

Perhaps it's all nonsense, *E pur si muove!*

Acknowledgements G. B. expresses his unbounded gratitude to Charles H. Bennett, who introduced him to the wonders of quantum mechanics more than 30 years ago. In a very real sense, he can trace back most of his beliefs to Bennett's patient preaching. He is also deeply grateful to Christopher Fuchs, his malt brother, with whom he has had most fascinating discussions concerning the foundations of quantum mechanics. The above acknowledgements should not be construed into saying that Bennett and Fuchs would agree with all that has been written here!

We also benefitted from several crucial discussions with too many wonderful people to list them here even though they all helped shape our beliefs, but we must mention at least Jeffrey Bub, Claude Crépeau, Patrick Hayden, Nicolas Gisin, Adrian Kent, Nathaniel David Mermin, the late Asher Peres, and Alain Tapp. Even though G. B. has lived with these ideas for several decades, and many people throughout the years have asked him if he had anything written about it, who knows how much longer it would have been before he wrote them down, if ever, without the opportunity and motivation provided by Antoine Suarez and his 2010 meeting in Barcelona on *Is Science Compatible with Our Desire for Freedom?*[15]

G. B. is supported in part by the Natural Sciences and Engineering Research Council of Canada, the Canada Research Chair program and the Canadian Institute for Advanced Research.

References

Aspect, A., Dalibard, J., Roger, G. (1982a). Experimental test of Bell's inequalities using time-varying analyzers. *Physical Review Letters*, *49*, 1804–1807.

Aspect, A., Grangier, P., Roger, G. (1981). Experimental tests of realistic local theories via Bell's theorem. *Physical Review Letters*, *47*, 460–463.

Aspect, A., Grangier, P., Roger, G. (1982b). Experimental realization of Einstein-Podolsky-Rosen-Bohm Gedankenexperiment: A new violation of Bell's inequalities. *Physical Review Letters*, *49*, 91–94.

Bell, J.S. (1964). On the Einstein-Podolsky-Rosen paradox. *Physics*, *1*, 195–200.

Bell, J.S. (1990). Against "measurement". *Physics World*, *3*(8), 33–40.

Blaylock, G. (2010). The EPR paradox, Bell's inequality, and the question of locality. *American Journal of Physics*, *78*, 111.

Byrne, P. (2007). The many worlds of Hugh Everett. *Scientific American, 297*(6), 98–105.

Cirel'son (Tsirelson), B.S. (1980). Quantum generalizations of Bell's inequality. *Letters in Mathematical Physics*, *4*, 93–100.

Clauser, J.F., Horne, M.A., Shimony, A., Holt, R.A. (1969). Proposed experiment to test local hidden-variable theories. *Physical Review Letters*, *23*(15), 880–884.

[15] http://www.socialtrendsinstitute.org/Activities/Bioethics/Is-Science-Compatible-with-Our-Desire-for-Freedom.axd, accessed 29 February 2012.

Deutsch, D. & Hayden, P. (2000). Information flow in entangled quantum systems. *Proceedings of the Royal Society of London, Series A, 456*(1999), 1759–1774. http://arxiv.org/abs/quant-ph/9906007.

Duhamel, V. & Raymond-Robichaud, P. (2011). Guildenstern and Rosencrantz in quantumland – A reply to Adrian Kent. http://arxiv.org/abs/1111.2563. Accessed on 10 Nov 2011.

Einstein, A., Born, M., Born, H. (1971). *The Born-Einstein letters: correspondence between Albert Einstein and Max and Hedwig Born from 1916 to 1955*. Walker & Company, New York.

Einstein, A., Podolsky, B., Rosen, N. (1935). Can quantum-mechanical description of physical reality be considered complete? *Physical Review*, *47*, 777–780.

Everett III, H. (1957). "Relative state" formulation of quantum mechanics. *Reviews of Modern Physics*, *29*(3), 454–462.

Freedman, S.J. & Clauser, J.F. (1972). Experimental test of local hidden-variable theories. *Physical Review Letters*, *28*, 938–941.

Gisin, N. (2013). Are there quantum effects coming from outside space–time? Nonlocality, free will and "no many-worlds". In A. Suarez & P. Adams (Eds.), *Is science compatible with free will? Exploring free will and consciousness in the light of quantum physics and neuroscience*. New York: Springer. Chapter 3.

d'Holbach, P.-H. Thiry, Baron (1770). *Système de la nature, ou Des loix du monde physique & du monde moral* (originally published under the name of J.-B. de Mirabaud).

Kent, A. (2010). One world versus many: The inadequacy of Everettian accounts of evolution, probability, and scientific confirmation. In S. Saunders, J. Barrett, A. Kent, & D. Wallace (Eds.), *Many worlds? Everett, quantum theory and reality*, Chapter 10. Oxford: Oxford University Press.

Laplace, P.-S. de (1814). *Essai philosophique sur les probabilités*. Paris: Courcier.

Mermin, N.D. (1981). Bringing home the atomic world: Quantum mysteries for anybody. *American Journal of Physics*, *49*, 940–943.

Mermin, N.D. (1994). Quantum mysteries refined. *American Journal of Physics*, *62*, 880–887.

Nielsen, M.A. & Chuang, I. (2000). *Quantum information and computation*. Cambridge: Cambridge University Press.

Peres, A. (1995). *Quantum theory: concepts and methods*. Dordrecht: Kluwer.

Popescu, S. & Rohrlich, D. (1994). Quantum nonlocality as an axiom. *Foundations of Physics*, *24*(3), 379–385.

Pusey, M.F., Barrett, J., & Rudolph, T. (2012). On the reality of the quantum state. *Nature Physics, 8*(6), 474–477.

Rubin, M.A. (2001). Locality in the Everett interpretation of Heisenberg-picture quantum mechanics. *Foundation of Physics Letters*, *14*(4), 301–322.

Salart, D., Baas, A., Branciard, C., Gisin, N., Zbinden, H. (2008). Testing the speed of 'spooky action at a distance'. *Nature*, *454*, 861–864.

Schrödinger, E. (1935). Discussion of probability relations between separated systems. *Mathematical Proceedings of the Cambridge Philosophical Society*, *31*(4), 555–563.

Stefanov, A., Zbinden, H., Gisin, N., Suarez, A. (2002). Quantum correlations with spacelike separated beam splitters in motion: Experimental test of multisimultaneity. *Physical Review Letters*, *88*(12), 120404.

Suarez, A. & Scarani, V. (1997). Does entanglement depend on the timing of the impacts at the beam-splitters? *Physics Letters A*, *232*(1–2), 9–14.

Chapter 5
Free Will and Nonlocality at Detection as Basic Principles of Quantum Physics

Antoine Suarez

Abstract Quantum physics highlights the crucial role of free will and nonlocality at detection as basic principles of science. There is interplay of free will and nonlocality: On the one hand, the assumption of free will is necessary to prove nonlocality; on the other, experiments demonstrate that nonlocal effects come from outside space-time, and thereby stress the nonmaterial character of free will. Additionally, recent experiments demonstrate that the basic principles ruling the material world, like the conservation of energy, require nonmaterial coordination. Finally, quantum "indeterminism" does not necessarily mean lack of purpose and control: Randomness is always accompanied by nonmaterial control, even at the level of quantum devices in labs.

Keywords Nonlocality at detection • Conservation of energy • Empty wave • Many worlds • Parallel lives • Wake–sleep cycle • Irreversibility • Free will • Limited consciousness

5.1 Introduction

The philosopher Immanuel Kant was fully aware that the principle of freedom conflicts with the deterministic description of classical physics as he declared: "it cannot be alleged that, instead of the laws of nature, laws of freedom may be introduced into the causality of the course of nature. For, if freedom were determined according to laws, it would be no longer freedom, but merely nature." (Kant 1787).

A. Suarez (✉)
Center for Quantum Philosophy, The Institute for Interdisciplinary Studies, Berninastr. 85, 8057 Zurich, Switzerland

Social Trends Institute/Bioethics, Barcelona, Spain
e-mail: suarez@leman.ch

A. Suarez and P. Adams (eds.), *Is Science Compatible with Free Will?: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience*,
DOI 10.1007/978-1-4614-5212-6_5,

Although he clearly saw the conflict between freedom and deterministic science, he decided to live with it, without giving up freedom and without questioning deterministic science. Rewording Henri Bergson one could say that Kant's unconscious belief in freedom was so unbreakable that he refused to sacrifice it on the altar of determinism; in order to save freedom Kant transferred it into the realm of the unintelligible things we cannot know (Bergson 1888).

Science and the principle of human freedom continue to co-exist today according to Kant's assumption of nonoverlapping realms. So, for instance, the German neuroscientist Wolf Singer states, "We experience ourselves as free mental beings, but the scientific view does not admit any room for a mental agent like free will, which influences neurons and produces actions [. . .]. In my eyes this conflict cannot be solved for the time being. Both descriptions can be shared even by researchers of the brain: when I observe the brain I cannot find any evidence of a mental agent like free will or personal responsibility—nevertheless when I get home in the evening I hold my children responsible for their actions if they have done any nonsense." (Singer 2000) Indeed, most working scientists seek, so to speak, to have it both ways. When talking to colleagues, they support the view that our brain functions according to deterministic laws, but in private conversation they declare, "I believe also in human freedom, but on the philosophical and moral level."

Yet the belief that human freedom and science occupy separate realms seems flawed. It founders on the fact that by assuming freedom one makes claims about the physical world that turn out to have scientific implications. So for instance, when I claim to be the author of this chapter and to be expressing original thoughts in it, I assume that, in writing the text, I control the firing of my neurons and thereby the movement of my fingers through the exercise of my own free will. That is, these movements are not *completely* predetermined at the beginning of the Universe, or even 1 month ago. Anyone who claims the right "to choose how to live his life" should, to be consistent, exclude any explanation of his brain using *only* deterministic causality, be it in terms of genes, chemicals, or environmental influences. The assumption that human behavior is not *completely* determined by the past in fact plays a key role in the way we behave in daily life and organize society through law.

One gets the impression that scientific achievements fascinated Kant and many other thinkers so much, that they didn't dare question determinism. However, to be consistent, their position should rather have been: Since I am not ready to renounce freedom, determinism cannot be the last word in science.

In fact, the new science arrived in the form of quantum physics. In Sect. 5.2 of this chapter, I argue that free will and quantum nonlocality enhance each other: the experimental demonstration of nonlocality assumes the free will of the experimenter to some extent, and the nonmaterial character of free will can be better understood in the light of nonlocal effects coming from outside space-time. Sect. 5.3 stresses the inseparability of material and nonmaterial principles: A recent experiment demonstrates that the conservation of energy requires nonmaterial (nonlocal) agency. Sect. 5.4 discusses the local models of "Empty waves," "Many worlds," and "Parallel lives." In the Sects. 5.5, 5.6 and 5.7 I argue that quantum randomness can in principle be controlled by free will and speculate how

this may happen in the brain: On the one hand the wake–sleep cycle appears to be a fundamental principle of science; on the other, the "wet and hot" neuronal environment is a positive feature for the quantum functioning of the brain. Sect. 5.8 discusses the relationship between consciousness, irreversibility, and the observation process. Sect. 5.9 highlights some conclusions.

5.2 The Interplay of Free Will and Quantum Nonlocality

A frequent objection to the possible relevance of quantum physics to the question of free will is that quantum nondeterministic randomness excludes the possibility of order and control, and therefore free will. Thus one states for instance that: "In the end, however, it is clear that neither determinism nor randomness is good for free will. If nature is fundamentally random, then the outcomes of our actions are also completely beyond our control: randomness is just as bad as determinism" (Vedral 2006). Here the term "randomness" refers to "exclusion of order and control", rather than "not completely determined by the past". Actually, this objection is a variation of Hume's argument that "the universe goes on for many ages in a continued succession of chaos and disorder" (Hume 1779): Hume's criticism of temporal causality leads him to assume indeterminism in the world (contrary to Kant) but he excludes the possibility that invisible nonmaterial principles rule the visible material world (similar to Kant).

However quantum physics does not demand the presumed incompatibility of quantum randomness with order and control. Quantum physics tell us rather that quantum randomness is inseparable from nonmaterial principles acting and controlling randomness from outside space-time.

In so-called entanglement experiments, correlated events appear in regions far away from each other. The separation is such that no material connection can explain the correlations. Things can be described appropriately as follows: In a lab (say in Zürich) a physicist has a device for detecting photons (light particles) with two detectors denoted D(0) and D(1). For each photon only one of the two detectors fires, that is, the measurement yields either the result 1 or 0, according to whether the detector D(1) or the detector D(0) fires. After many runs the results obtained (1,0,0,1,0...) are distributed like the results one gets by tossing a fair coin (with 1 for head and 0 for tail), that is, they do not exhibit any particular pattern or order. In another lab far away (say New York) another physicist with a similar device gets similar results, that is 1s and 0s distributed like the heads and tails of tossing a coin. Random results in Zürich and random results in New York. The results in Zürich and New York occur pair-wise: for each result in Zürich there is a corresponding result in New York happening almost simultaneously. Because of the distance and insignificant time interval between the detection in Zürich and that in New York, the two results cannot be connected by any signal traveling at the speed of light. Nonetheless when the physicists come together and compare their results they see that there is perfect correlation: when the result in New York is 0, the result in Zürich is 0, and when the

result in New York is 1, the result in Zürich is 1. Randomness in Zürich, and randomness in New York, but the same randomness in both places!

By the magnificent mathematical result obtained by John Bell in 1964 it is possible to exclude the possibility that the photons behave like genetic twins determined by "hidden genetic programs" that explain the correlations. This possibility is excluded using more sophisticated experiments, in which each experimenter has the choice of switching his device into two different configurations. Depending on the experimenters' choices the results in New York and Zürich are correlated, but not perfectly. Bell's theorem imposes a limit to the degree of correlation hidden programs can achieve in these types of experiments (Bell 1987). The degree of correlation experimentally observed and also predicted by quantum theory violates Bell's limit. This means that one cannot account for these correlations by means of common causes in the past, unless each experimenter himself is predetermined in choosing the settings of his apparatus (Gisin 2013).

Hence, *if one assumes that the experimenters are free to some extent*, the correlations cannot be explained either by a common cause in the past or by any material signal traveling in space-time from one place to the other.

Even more sophisticated experiments involving devices in motion (relativistic experiments) demonstrate that these quantum correlations are independent of any temporal order (before–before experiment), and in this sense come from outside space-time (Stefanov et al. 2002, 2003). This has now been confirmed by a theoretical result as well (Bancal et al. 2011).

What does it mean to say that these quantum correlations come from outside space-time? Nicolas Gisin answers that such correlations cannot be explained by any story encoded in space-time (Gisin 2013). This is another way of saying that the agency behind the correlations cannot be *directly* accessed by any detection or observation. Space-time is the realm of material, observable things. To be in space-time means to be accessible to experimental control.

Nonlocality experiments support the view that causation in time is an illusion of our intuition. The true causes are invisible and act from outside space-time. We know these causes indirectly through the visible effects they produce, like for instance nonlocal correlations.

This result allows us to consider free will as a nonmaterial capability of the experimenter as well. By means of his free will the experimenter controls the dynamics of his brain, from outside space-time. In this sense one can say that free will is a spiritual principle.

In summary, on the one hand, for deriving nonlocality it is necessary to assume "free will" (at least to a certain extent), that is, the experimenter can freely choose at least some parameters of the setup. But on the other hand, nonlocality shows that free will can be considered to be a nonmaterial principle, that is, a capability bringing about observable effects that cannot be explained exclusively by stories stored in space-time (material causes).

Nonetheless, as we will see in the following Sects. 5.3 and 5.4, there is another type of nonlocality that is more basic than Bell's nonlocality.

5.3 The Material World Emerges from Nonmaterial Features

Actually nonlocality appears already in the most basic quantum phenomenon of interference and it is also important to have conservation of the energy in each individual quantum process (and not only on the average). This is demonstrated by a recent experiment.

Consider the quantum interference experiment sketched in Fig. 5.1a. When one works with a sufficiently weak intensity of light, then only one of the two detectors clicks: either D(0) or D(1). Things happen as if light consists of packets of energy called photons: The energy of each photon is determined by the light's frequency (photoelectric effect), and the detections obey the rule: "one photon, one count."

Nevertheless, for calculating the counting rates of each detector one must take into account information about the two paths leading from the laser source to the detectors (interference effect): The probability that one detector clicks in a large series of runs can be exactly predicted for each detector and depends deterministically on the optical path length difference (as represented in Fig. 5.1a).

According to *standard* quantum mechanics, which detector clicks in a single run (the outcome) is decided by a true choice (on the part of Nature) at the moment of

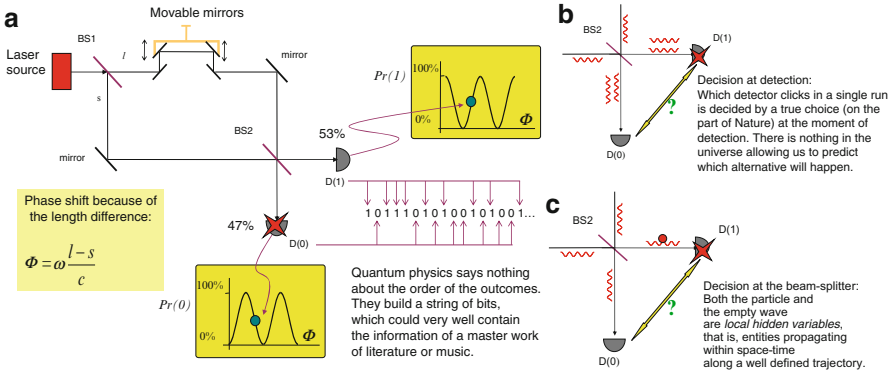


Fig. 5.1 (**a**) Laser light of frequency ω emitted by the source is either transmitted or reflected at each of the beam-splitters (half-silvered mirrors) BS1 and BS2; therefore the light can reach the detectors D(1) and D(0) by the paths l and s; the path-length l can be changed by the experimenter. With a sufficiently weak intensity of light, only one of the two detectors clicks: either D(1) or D(0) (photoelectric effect). Nevertheless, Nature calculates the counting rates of each detector *Pr(1)* and *Pr(0)* taking account of the length of the two paths l and s (interference effect). (**b**) According to standard quantum mechanics, which detector clicks in a single run is decided by a true choice (on the part of Nature) when the information about the two paths reaches the detectors. (**c**) According to Louis de Broglie's picture the outcome becomes determined at the beam-splitter: The particle leaves by one output port, and an "empty wave" (without energy and momentum) leaves by the other. This wave is inaccessible to observation but influences the particle when both meet together and thereby ensures the possibility of interference

detection, taking account of the information about the two paths (Fig. 5.1b). In other words, interference requires coordination at detection to determine whether either D(0) or D(1) fires.

The *standard* view has the following noteworthy implication: Since D(0) and D(1) can be arbitrarily far away from each other (Fig. 5.1b), the coordination between these two detectors cannot be explained by local causality, that is, signals propagating between the detectors with velocity $v \leq c$ (energy propagation is upper bounded by the velocity of light c). Thus, the choice that one detector does click and the other doesn't click cannot be explained by material causality, by any story or law recorded in space-time. In this sense, the decision at detection is nonlocal and comes from outside space-time.

Implicit in this conclusion is the assumption that the experimenter is free to choose the path length after the photon leaves the beam-splitter BS1 (that is, the experimenter is not determined by the path the photon takes) and the experimenter's choice does not determine (backwards in time) which path the photon takes.

The assumption of nonlocality at detection was already implicit in the idea of the *wave function collapse* (also called Copenhagen interpretation), and raised Einstein's suspicions already in the Fifth Solvay Conference (1927), years before the celebrated EPR argument (1935). So, historically, *nonlocality at detection* appears before *Bell's nonlocality* and begets it to some extent (Suarez 2012).

Suppose an experiment with detectors D(0) and D(1) near to each other and a counting rate of 50% in each detector. Suppose that one wishes to maintain both the local explanation of things by means of signals traveling at velocity less than or equal to that of light, and the *standard* view of decision at detection (shared by most physicists). According to this *local theory*, if one puts the detectors D(0) and D(1) distant enough from each other, then they remain uncoordinated and will fire randomly. Consequently, for each single photon one will have two counts in 25% of the events (one count in each detector), no count in 25% of the events, and one count (in either D(1) or D(0)) in 50% of the events. This means: energy is conserved on average, but not in each individual quantum process.

These predictions have been tested and ruled out by a recent experiment (Guerreiro et al. 2012). The reported falsification of the *local theory* means that (as far as one assumes decision of the outcome at detection) the conservation of energy in individual quantum processes is inseparably related to nonlocality. This means that the most fundamental principle ruling the material world, the conservation of energy, requires nonmaterial coordination of detection outcomes. Additionally, the experiment is a natural and most direct demonstration of nonlocality in a context where the violation of Bell inequalities cannot be used as a criterion for establishing nonlocality. In this sense, the experiment highlights the fact that the principle of nonlocality rules the whole of quantum physics: it emerges already in interference phenomena involving only two detectors and not only when four or more detectors are involved, as in Bell type experiments.

5.4 "Empty Waves," "Many Worlds," "Parallel Lives," and Nonlocality at Detection

The result presented in Sect. 5.3 above could be questioned simply by assuming that the outcomes become determined at the beam-splitters (Fig. 5.1c: "decision at the beam-splitter"). Then detections on one output port of a beam-splitter do not influence detections on the other output port, and it is possible to escape "nonlocality at detection." However, local models assuming the beam-splitters as the devices where the outcome's choice happens, necessarily involve local hidden variables of the de Broglie's "empty pilot wave" type, that is, the assumption that after leaving the beam-splitter the particle always follows a well-defined path and an "empty" wave (inaccessible to observation) follows the alternative path (Fig. 5.1c).

Actually, the "local models" addressed in the conventional Bell experiments are "local pilot wave models." When implemented in entanglement experiments they imply correlated outcome decisions at two space-like separated beam-splitters. Since "local pilot wave models" fulfill the well-known locality criteria of Bell inequalities, they are refuted by the experimental violation of such inequalities (Bell 1987). John Bell himself emphasized that the true "hidden variable" is the "pilot wave." Accordingly violation of Bell inequalities implies that the decisions at two beam-splitters are nonlocally correlated, even if one assumes local "empty waves."

It is worth noting that "empty waves" as *local* hidden variables are unobservable and inaccessible in principle. Thus, the very concept of "empty wave" is somewhat logically inconsistent inasmuch as it refers to an entity existing and propagating locally in space–time (within the light cone), but which one cannot detect or control directly.

De Broglie's idea was further elaborated by David Bohm, who added a "nonlocal quantum potential" to the empty wave. Bohm's nonlocal model is compatible with the violation of Bell inequalities and accounts for the nonlocal quantum correlations in entanglement experiments. Indeed, it was the nonlocal feature of Bohm's model that inspired Bell in the search for his inequality (Bell 1987, p. 128).

However, Bohm's picture has a "romantic counterpart" (as John Bell said), which is able to restore locality in entanglement experiments: The "many worlds interpretation" (Bell 1987 p. 192). The basic idea of this picture is that at the beam-splitter the world W gets split into two different worlds W' and W": In world W' the particle takes path "s" and the empty wave path "l", and in world W" the particle takes path "l" and the empty wave path "s". So there is no choice but all possible alternatives get realized, although in different worlds which thereafter cannot interact with each other.

Regarding "many worlds" John Bell stated:

> "The 'many world interpretation' seems to me an extravagant, and above all an extravagantly vague, hypothesis. I could also dismiss it as silly. And yet. . . It may have something distinctive to say in connection with the "Einstein Podolsky Rosen puzzle," and it would be worthwhile, I think, to formulate some precise version of it to see if this is really so." (Bell 1987, p. 194).

Work by Lev Vaidman (2002) and more recently by Gilles Brassard and Paul Raymond-Robichaud (2013) show that "many worlds" has really something distinctive to say in connection with the "EPR puzzle": Refutation of local hidden

variables (by the violation of Bell's inequalities or other means) doesn't mean refutation of locality.

This idea is elaborated by Gilles Brassard and Paul Raymond-Robichaud in this volume as the theory of "parallel lives" (2013), which they also describe as a belief characteristic of a scientific community called "the Strong Church of the Larger Hilbert Space." According to "parallel lives" observers and their apparatuses split. So for instance Alice with her apparatus lives inside a bubble and when she performs a measurement the bubble splits into two bubbles. Inside one bubble Alice sees one outcome; and inside the other Alice's copy sees the alternative outcome: "From now on, the two bubbles are living parallel lives. They cannot interact between themselves in any way and will never meet again." These authors argue that the theory of "parallel lives" reconcile violation of Bell inequalities with "a fully deterministic, strongly local and strongly realistic interpretation of quantum mechanics" (Brassard and Raymond-Robichaud 2013).

The theory of "parallel lives" has the merit of highlighting this important point: if one assumes that the decision of the outcomes happens at the beam-splitter BS2 in the experiment of Fig. 5.1a, then one has to accept the "empty wave" (Fig. 5.1c) and, in the end, one is led to local interpretations of quantum mechanics like "many worlds" and "parallel lives," which are compatible with the violation of Bell inequalities.

I would like to stress that assuming "free will" (Gisin 2013) is not sufficient to reject "many worlds" or "parallel lives": one has to reject "empty waves" as well, and therefore accept that the decision of the outcome happens at detection. Otherwise one is led to contradictions.

Indeed if one accepts nonlocality, the rejection of decision of outcomes at detection is obviously no longer motivated by the wish to escape nonlocality. Thus, by assuming both nonlocality and decision at the beam-splitter, one implicitly accepts determinism, i.e., that correlated outputs of devices are necessarily determined by some cause in the past light-cone, even when these devices are far away from each other. Therefore one must with even better reason assume that the outputs of the experimenter's brain are predetermined as well, that is, it does not make sense to assume that the experimenter is free to choose the input values of his/her apparatus. But by discarding the experimenter's freedom one gets rid of nonlocality as well.

In summary, if one assumes decision of outcomes at the beam-splitter one can neither have experimenter's freedom nor prove nonlocality. By contrast, if one assumes the decision of the outcomes happens at detection, then one can have experimenter's freedom and experimental demonstration of nonlocality (Guerreiro et al. 2012, Suarez 2012).

Are there reasons allowing us to prefer one assumption to the other? Surely this cannot be decided by experiment. However, I see two strong reasons in favor of the standard view of decision at detection:

– *Scientific consistency*:

We are being taught by quantum physics that science is based on the following two principles:

Principle A: All that is in space-time is accessible to observation (except in case of space-like separation).
Principle Q: Not all that matters for physical phenomena is contained in space-time.

By assuming decision at the beam splitter you become a "strong Faithful" of the "Church of the Larger Hilbert Space" (Brassard and Raymond-Robichaud 2013), because (without even realizing it) you profess rejection of these two principles A and Q, and this rejection is the main article of the "many worlds" faith. Hence you will not have the necessary mental strength to reject locality.

By contrast, if you assume nonlocal decision at detection you profess both principles A and Q, and, therefore, you remain outside the Church of "many worlds" and will be able to oppose this interpretation without contradicting yourself. And you can also consistently reject superdeterminism for the sake of freedom.

"Empty waves," "many worlds," and "parallel lives" reject both principles, and in particular deny that the only way to have inaccessibility within space-time is through space-like separation (*Principle A*). I think that this principle is a very reasonable way of defining space-time, and it should be assumed by nonlocal quantum mechanics as well (and in fact the orthodox interpretation assumes it). Einstein himself apparently never got rid of this principle, and this may be the reason why he never definitely endorsed de Broglie's "empty waves."

Therefore, for reasons of scientific consistency one should reject "empty waves," "many worlds," and "parallel lives." And in any case one cannot say that "many worlds" reconciles quantum mechanics with Einstein's local realism because it is at odds with both.

Notice that in accordance with the principle that space-like separation is the only way to make entities or regions in space-time inaccessible, we assume that the nonlocal outcomes at detection come from outside space-time just because their origin is inaccessible in principle.

In summary, probably in accord with Einstein but in conflict with "empty waves" and "many worlds," we assume that the human observer can in principle access all what lies in space-time provided it is not space-like separated, and in conflict with both (Einstein and "many worlds"), we assume nonlocal decision at detection coming from outside space-time, and therefore the human observer cannot in principle access all that matters for the physical reality.

– *Free will, personal identity and authorship*

As far as one assumes nonlocal coordinated outcomes at detection, I share the arguments Nicolas Gisin gives in favor of free will against "many worlds" (see Gisin 2013).

In addition, I would like to stress that by assuming personal identity and authorship (crucial for granting personal rights) one is assuming agency coming from outside space-time. Anyone claiming to have a personal identity implicitly accepts that his or her identity is conserved in time. As the author of this chapter, for example, I claim to be the same person as the author of other papers I wrote years ago. That is, my personal identity has roots outside time. In this sense I, and the paper I am writing now, cannot be explained exclusively by material or observable causal chains. If someone admits only material causal chains within time (even more in "many worlds" and "parallel

lives"), he will deny that personal identity persists through time. Personal identity vanishes as an illusion if the "self" is not somewhat outside time.

Interestingly, Gilles Brassard at the end of his article asks:

> "It may be that free will can at best be an illusion in a world ruled by the Strong Church of the Larger Hilbert Space because every time you think that you make a decision [. . .], you also make the complementary decision in the Universal superposition. However, what does it matter if free will does not truly exist, provided the illusion is perfect?"

And in a note remarks:

> "Seriously, we would not want to live in the Matrix imagined by the Wachowski siblings, no matter how perfect is the simulation. So, perhaps we do care after all!" (Brassard and Raymond-Robichaud 2013, note 14, p. 59)

These words eloquently illustrate that the ultimate reason for choosing free will may be the profound desire of ensuring personal identity and authorship, and so making it possible to claim personal rights. Following Kant one could say that freedom of will must be presupposed as a quality of all rational beings (Kant 1785). No matter how intellectually gratifying a world picture may be in which the wavefunction collapses unitarily (Brassard and Raymond-Robichaud 2013), personally I prefer to have a science that allows me to defend my rights, and I am only interested in discussing with people who are interested in defending their rights in a coherent manner.

5.5 Quantum "Randomness" can be Controlled by Free Will

It is important to note that one cannot account for nonlocal effects by invoking "indeterminism" or "randomness" (in the sense of lack of control) alone. In quantum interference and entanglement phenomena, randomness and control appear inseparably united in the same phenomenon. There is randomness in Zürich and randomness in New York, but controlled by some agency producing "the same randomness" in separated regions.

The event that D(0) clicks and D(1) does not click is often said to be "a genuinely random event." However I dare to insist, here "random" does not necessarily mean "lack of control" but rather that Nature's decision about which detector clicks, though it has some roots in the past, is not completely determined by the past. Which detector clicks is not only unpredictable by us because we don't yet know the formula connecting the past with the present and the present with the future, but it is unpredictable in principle because such a formula doesn't exist at all.

A single outcome, either "0" or "1", represents a bit of information, and a series of outcomes (a bit string) builds a piece of information. Quantum physics requires that long series of outcomes fulfill a statistical distribution imposed mainly by the parameter of the path-length difference. Quantum physics imposes nothing regarding the order in which the single outcomes occur, and neither establishes how long a series must be, to be considered "long."

Thus, it is possible in principle that an unobservable mental variable (a free-willed intellect) deliberately influences the order of the bits for a time and encodes a message

in the string. The result that an immaterial agency can control quantum randomness may help to explain how a mind purposefully generates information.

If by "random" one means events that are not completely determined by the past, one can very well consider that quantum randomness and free will have the same origin. Your free will is for me as unpredictable as the best random number generator. One can interpret the before–before experiment (Stefanov et al. 2002, 2003) in terms of "nonlocal randomness" or "nonlocal free will" as well. Nicolas Gisin states: "the same randomness manifests itself at several locations." I state: "A mind influences local randomness from outside space-time to produce nonlocal order." I think these two statements are equivalent.

In summary, metaphysics is based on observation, and today's physics provides experiments that may give rise to metaphysical reflection. While quantum randomness allows for freedom of action, it does not exclude control over our actions. To this extent, it seems that free will is not calling for a new physics.

5.6 Brains as Quantum Devices

To this quantum philosophical view one could object that in the lab we do not meet quantum devices printing out poems, scientific papers, or any meaningful message.

I would reply that such a possibility is not forbidden at all by quantum physics. What is more, in daily life we meet plenty of "quantum devices" (human brains) communicating efficiently with each other. No human being can manipulate a quantum interferometer through material agency and oblige it to print out a determined piece of information. However, a human being can steer the order of the outcomes in his brain through nonmaterial agency and produce a meaningful piece of information: a talk, a paper, a masterpiece of literature, or music.

There is nothing against the assumption that the neural activity responsible for the chapter I am writing now cannot be explained exclusively by material or observable causal chains, but requires spiritual agency from outside space-time. Although the neural activity responsible for the spontaneous movements of the human body is a visible effect, it cannot be explained exclusively by a chain of visible causes, by pure material deterministic agency. The choices guiding my spontaneous movements, for instance typing a particular key ("r", "a", "d", "o", etc.) while writing this paper, originate from an unobservable mental agency influencing the basic random dynamics of my brain.

But now one could object the other way around: If simple devices like beam-splitters and interferometers can in principle be controlled by a free willed intellect, "why should an autonomous and self-conscious mind need a brain to live and act in the material world?" (Roth 1997, 1999).

To this objection I would answer that the human condition is defined by certain fundamental limitations. One of these is the *impossibility of being in a permanent conscious state*.

This limit seems to come from the fact that a human person is not a pure spiritual intellect (sort of "angel"), but a neuronal one, i.e., a mind that cannot be permanently conscious. A human brain is nothing other than the ensemble of conditions which make it possible that such a mind exists, basically through a wake–sleep cycle. The human mind cannot be permanently aware of its own existence, the human will cannot act on purpose all the time. My capacity for alertness, for instance, is limited: I cannot keep driving a car indefinitely without sleeping; after a time I will begin to have random neuron firings, eventually hallucinate, and finally fall asleep at the wheel.

Astonishingly, sleep is still a poorly understood phenomenon, and nobody knows what constants of nature determine the wake–sleep cycle. Our view is that this cycle is a very fundamental fact, which determines the physics of the world we live in to a large extent.

Consider a brain that for some reason is incapable of producing purposeful behavior. One could compare such a brain to the interferometer of Fig. 5.1a producing outcomes with a very low level of nonmaterial (nonlocal) control, randomly distributed according to the statistics imposed by the physical parameters (path-length difference). Many features of my brain's physiology are susceptible to deterministic description in terms of observable causal chains (the metabolism involved in the arousal potentials triggering bodily movements, for instance, follows the usual physical conservation laws). Additionally, the physiological measurable parameters of the brain are fixed by a number of factors (genetic, epigenetic, and environmental ones), and in particular by signals coming from the senses. These parameters (like the path-length difference in an interferometer) characterize the statistical distribution of the brain outcomes when these happen without purpose, as is for instance the case with the eyes' movements during sleep, or the unconscious spontaneous behavior of hydranencephalic children or patients in persistent vegetative state (PVS patients).

Suppose now that a mind could control purposefully during a short period of time the outcomes of the interferometer in Fig. 5.1a, but after this period, for some reason, the mind continues producing purposeless outcomes. In the long term the outcome distribution will tend to the distribution corresponding to the physical parameters. In this very sense, during certain periods of time a brain produces meaningful pieces of information (speech, text, musical composition, painting, etc.). And during other periods of time (while sleeping and even during many waking periods) the brain produces uncontrolled random signals, which tend to restore the statistical outcome distribution the physiological parameters impose, and would in fact restore it if they would last forever.

In summary, there is no such a thing as "pure randomness," randomness without any control, in the world. Already at the most elemental level of quantum devices in the lab (like interferometers and random number generators) randomness is always accompanied by nonlocal control, that is, nonmaterial coordination of the detectors by means of influences coming from outside space-time. But there is no such a thing as pure consciousness in the world either. At the highest level of the human brain the wake–sleep cycle causes free will and consciousness to be inseparably united to periods of random behavior.

5.7 The Interplay of Coherence and Decoherence in Living Systems

It is well known that the brain provides a wet and hot environment where it is quite impossible to isolate conveniently a quantum system in order to get entanglement. This feature is often advanced as an objection against invoking the brain as an appropriate quantum interface for mental agency. However, the fact that the neuronal environment does not offer the conditions required to avoid decoherence may be not a problem but rather a welcome feature of the human brain for the aim of implementing the operations of free will and consciousness.

On the one hand, in the preceding Sect. 5.6 we have suggested that quantum interference (which is much more resistant to decoherence than entanglement) may be a better interface than entanglement for free will and consciousness. On the other hand, we are discovering today that biological phenomena require the interplay of quantum coherence and decoherence (Ball 2011). This interplay seems to be the reason for the high efficiency of phenomena like photosynthesis (Collini et al. 2010, Giorda et al. 2011, King et al. 2012). There are attempts also to use this interplay to explain the functioning of ion channels (Vaziri and Plenio 2010).

5.8 Free will, Consciousness, Irreversibility

Invoking quantum physics to explain mental operations often meets the objection that consciousness, thinking, and deciding are brain processes involving billions of neurons, i.e., macroscopic physical states, and therefore far away from quantum states (Roth 1997).

In the same line of thinking one states: "Molecular machines, such as the light-amplifying components of photoreceptors, pre- and postsynaptic receptors and the voltage- and ligand-gated channel proteins that span cellular membranes and underpin neuronal excitability, are so large that they can be treated as classical objects. Although brains obey quantum mechanics, they do not seem to exploit any of its special features." (Koch and Hepp 2006).

Such objections originate from a widespread prejudice about the quantum. One overlooks that the decision about which detector clicks (in an interference experiment, like the one represented in Fig. 5.1) does not happen when "one photon encounters a detector" but only subsequently, after a *virtual* cascade involving billions of electrons has been triggered. Only then an irreversible registration of a result happens and a human observer can become aware of it. In fact this means that the decision is not between "one photon encountering D(1)" and "one photon encountering D(0)," but between "a virtual assembly of electrons in D(1)" and "a virtual assembly of electrons in D(0)." The decision gives reality to one of these two virtual assemblies of electrons: detection is an "elementary act of creation," in John Wheeler's words.

The particular conditions defining *when* precisely the decision happens are to date an unsolved (but solvable) problem (the so-called measurement problem). By contrast, as said above, the question of which detector clicks is a matter of an unobservable free decision, and as such cannot be answered before the detection happens. All this means that "quantum effects" (already at the level of simple interference experiments) consist in decisions about macroscopic outcomes occurring in visible classical objects (detectors).

Let us now consider a conscious decision about, for instance, moving the right or left hand. We know that this depends on the building of different "transient neuronal assemblies" (Greenfield 2000). The neuronal assemblies (like the counts in different detectors in quantum experiments) are something macroscopic and measurable. But the choice between two rival neuronal assemblies, as the choice between two rival detectors, may very well originate from unobservable agency.

You can say that the difficulty in tackling this issue by experimental means is only due to the high inaccuracy of current measuring techniques: imaging techniques, for instance, are still too slow to capture the recruitment of ten million cells in less than a quarter of a second.

This was the kind of objection raised against indeterminism at the beginnings of quantum mechanics. The techniques for studying the neuronal activity will certainly improve, but not to the extent of overcoming any indeterminacy (Tononi 2013). Just as information technology will not improve to the extent of communicating faster than light.

If one assumes that basically "brains obey quantum mechanics," and you are for freedom, then the reasonable attitude is to conclude that the realization of one specific neuronal assembly among several possible ones cannot be explained *exclusively* through deterministic causality. As in the case of the detectors in the interference experiments, there is nothing in the observable universe, no story in space-time, capable of explaining why this neuronal assembly happens and not another (Tononi 2013).

As in the case of the detectors, the particular conditions defining *when* precisely a conscious decision happens are to date an unsolved (but solvable) problem. In both cases, detection and consciousness, the problem is related to the question of *irreversibility*: Nobody knows to date where it comes from.

Interestingly, the concept of *irreversibility* appears explicitly in the clinical definition of death, which basically states that death occurs when the neural functions responsible for certain spontaneous movements *irreversibly* break down. In establishing death this way, we are assuming as obvious that our capacity of restoring neuronal dynamics (our capability of reversing a process of decay) is limited in principle, even if we don't yet know where this limitation comes from. Similarly, one could assume that amplification in a photomultiplier and in the brain becomes irreversible in principle at a certain level, if beyond this level an operation exceeding the human capabilities would be required to restore the original quantum state. When such a level is reached the detector clicks, and consciousness arises.

To prevent a possible misunderstanding I would like to stress that in relating consciousness to measurement I am not endorsing the postulate that "consciousness

is strictly necessary to the collapse of the wave function." (This postulate is at the origin of the Schrödinger cat's paradox and has often been the object of criticism, for instance recently by Koch and Hepp 2006.) In fact, for measurement to happen it is not necessary at all that a human observer (conscious or not) is watching the apparatus. In a sense I consider the "collapse" to be something as objective as "death": For someone to die (generally) it is not necessary to be watched by some conscious physician. However the very definition of measurement makes relation to human consciousness: An event is "measured," i.e., *irreversibly* registered, only if it is possible for a human observer to become aware of it. Such a view combines the subjective and the objective interpretation of measurement: on the one hand no human observer has to be actually present in order that a registration takes place just the same as in the GRW "spontaneous collapse" (Bassi and Ghirardi 2007) or Penrose's "objective reduction" (OR) (Marshall et al 2003); on the other hand one defines the "collapse" or "reduction" with relation to the capabilities of the human observer (a near point of view is adopted by d'Espagnat 2006).

Even if measurement is basic to quantum mechanics, for the moment the theory does not define the conditions determining when the outcome gets *irreversibly* registered and measurement happens. This state of affairs clearly shows a point where quantum theory, as we know it today, can and must be completed. And to do it, it may be that we have to understand better how consciousness and free will happen in the brain. I am convinced that the solution of this problem will bestow on us a theory more fundamental than quantum mechanics.

5.9 Conclusion

We have seen that quantum physics stresses the following two principles as distinctive of the scientific attitude:

> *Principle A*: All that is in space-time is accessible to observation (except in the case of space-like separation).
> *Principle Q*: Not all that matters for physical phenomena is contained in space–time.

On this basis one can safely say that free will is perfectly compatible with today's science. By contrast, pictures rejecting these two principles, like "many worlds," are rather opposed to the foundations of science.

But there is more: Quantum physics is assuming free will as a basic axiom of science, together with the conservation of energy and nonlocal decision at detection. And improving our understanding of how free will and consciousness happen in the brain may be a promising road to improve quantum physics itself.

Acknowledgments I am thankful to Antonio Acín, Gilles Brassard, Nicolas Gisin, Bruno Sanguinetti, Valerio Scarani, Juleon Schins, Stefan Wolf, and Hugo Zbinden for stimulating discussions.

References

Ball, P. (2011). Physics of life: The dawn of quantum biology. *Nature, 474*, 272–274.

Bancal, J. D., Pironio, S., Acin, A., Liang, Y. C., Scarani, V., & Gisin N (2011) Quantum nonlocality based on finite-speed causal influences leads to superluminal signaling. *Nature Physics*, in press 2012, http://arxiv.org/pdf/1110.3795

Bassi, A., & Ghirardi, G. C. (2007). The conway-kochen argument and relativistic GRW models. *Foundations of Physics, 37*, 169–185.

Bell, J. S. (1987). *Speakable and unspeakable in quantum mechanics*. Cambridge: University Press.

Bergson, H. (1888). *Essai sur les données immédiates de la conscience: Conclusion*. http://classiques.uqac.ca/classiques/bergson_henri/essai_conscience_immediate/essai_conscience.pdf p. 102, Cited 5 May 2012

Brassard, G., & Raymond-Robichaud, P. (2013). Can free will emerge from determinism in quantum theory? In A. Suarez & P. Adams (Eds.), *Is science compatible with free will? (Chapter 4)*. New York: Springer.

Collini, E., Wong, C. Y., Wilk, K. E., Curmi, P. M. G., Brumer, P., & Scholes, G. D. (2010). Coherently wired light-harvesting in photosynthetic marine algae at ambient temperature. *Nature, 463*, 644–647.

d'Espagnat B. (2006). *On Physics and Philosophy*, Princeton and Oxford: Princeton University Press, Chapter 4.

Giorda, P., Garnerone, S., Zanardi, P., Lloyd, S. (2011). *Interplay between coherence and decoherence in LHCII photosynthetic complex*. http://arxiv.org/abs/1106.1986v1 Cited 5 May 2012

Gisin, N. (2013). Are there effects coming from outside space-time? Nonlocality, free will and "no many-worlds". In A. Suarez & P. Adams (Eds.), *Is science compatible with free will? (Chapter 3)*. New York: Springer.

Greenfield, S. (2000). *The Private Life of the Brain*. London: Penguin Books.

Guerreiro, T., Sanguinetti, B., Zbinden, H., Gisin, N., & Suarez, A. (2012). Single-photon space-like antibunching. *Phys Let A*, *376*: 2174–2177. http://dx.doi.org/10.1016/j.physleta.2012.05.019

Hume, D. (1779). Dialogues Concerning Natural Religion, Part 8. Project Gutenberg http://www.gutenberg.org/files/4583/4583-h/4583-h.htm#chap08 Cited 19 March 2012

Kant, I. (1785). Grundlegung zur Metaphysik der Sitten, 3. Abschnitt. http://gutenberg.spiegel.de/buch/3510/1, Groundwork for the Metaphysic of Morals, Chapter 3. http://www.earlymoderntexts.com/f_kant.html. Cited 19 September 2012.

Kant, I. (1787). Kritik der reinen Vernunft, 2. Auflage, Kapitel 94 http://gutenberg.spiegel.de/buch/3502/94 The Critique of pure reason, Second edition. http://books.google.ch/books?id=21cbrXR9JZUC&pg=PA219&lpg#v=onepage&q&f=false Cited 17. September 2012.

King, C., Barbiellini, B., Moser, D., & Renugopalakrishnan, V. (2012). Exactly soluble model of resonant energy transfer between molecules. *Physical Review B, 85*, 125106.

Koch, C., & Hepp, K. (2006). Quantum mechanics in the brain. *Nature, 440*, 611–612.

Marshall, W., Simon, C., Penrose, R., & Bouwmeester, D. (2003). *Physical Review Letters, 91*, 130401.

Roth, G. (1997). *Das Gehirn und seine Wirklichkeit*. Frankfurt: Suhrkamp. p. 284, 300-311.

Roth, G. (1999). Interview in *Morgenwelt*, 02-1999.

Singer, W. (2000). Wer deutet die Welt, Die Zeit, 50/2000, p. 3 http://www.zeit.de/2000/50/index Cited 5 May 2012

Stefanov, A., Zbinden, H., Gisin, N., & Suarez, A. (2002). Quantum correlations with spacelike separated beam splitters in motion: experimental test of multisimultaneity. *Physical Review Letters, 88*, 120404.

Stefanov, A., Zbinden, H., Gisin, N., & Suarez, A. (2003). Quantum entanglement with acousto-optic modulators: 2-photon beatings and Bell experiments with moving beamsplitters. *Physical Review A, 67*, 042115.

Suarez, A. (2012). "Empty waves", "many worlds", "parallel lives", and nonlocal decision at detection. http://arxiv.org/abs/1204.1732 Decission at the beam-splitter, or decision at detection, that is the question. http://arxiv.org/abs/1204.5848

Tononi, G. (2013). On the irreducibility of consciousness and its relevance to free will. In A. Suarez & P. Adams (Eds.), *Is science compatible with free will? (Chapter 11)*. New York: Springer.

Vaidman, L. (2002). The Many-Worlds Interpretation of Quantum Mechanics. In: Zalta EN (ed.) *The Stanford Encyclopedia of Philosophy* (Summer 2002 Edition).

Vaziri, A., & Plenio, M. B. (2010). Quantum coherence in ion channels: resonances, transport and verification. *New Journal of Physics, 12*, 085001.

Vedral, V. (2006). Is the universe deterministic? *New Scientist, 18*, 55.

Chapter 6
Are Humans the Only Free Agents in the Universe?

Zeeya Merali

Abstract In 2006, John Conway and Simon Kochen published their provocatively titled Free Will Theorem. The theorem, it is claimed, proves that if humans are truly free agents, then so too are elementary particles. As such, it strikes a blow against the suite of deterministic models of physics proposed as alternatives to the sstandard form of quantum mechanics. Here, I outline Conway and Kochen's proof and discuss some criticisms that have been laid against the philosophical claims made by its authors. I also assess the implications of the Free Will Theorem for physics and for the source of human free will, in light of recent work by philosophers and physicists aiming to incorporate elements of quantum mechanics into libertarian models of free will.

Keywords Quantum mechanics • Indeterminism • Free will • Libertarianism • Determinism

6.1 Introduction

> "Again, if all motion is always one long chain, and new motion always arises out of the old in order invariable, and if the first-beginnings do not make by swerving a beginning of motion such as to break the decrees of fate, that cause may not follow cause to infinity, whence comes this free will in living creatures all over the earth. . .?"
>
> Lucretius (circa 50 BCE)

> "We must believe in free will—we have no choice."
>
> Isaac Bashevis Singer (Kanfer 1997)

Z. Merali (✉)
Foundational Questions Institute, PO Box 3655, Decatur, GA 30031, USA
e-mail: merali@fqxi.org

A. Suarez and P. Adams (eds.), *Is Science Compatible with Free Will?: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience*,
DOI 10.1007/978-1-4614-5212-6_6,

Was I predestined to write this paper? Mathematicians John Conway and Simon Kochen would argue that I was not, since they are firm believers in the notion that humans are free agents, whose choices and actions are not entirely predetermined by events in the past history of the universe. In 2006, they went a step further by publishing their provocative *Free Will Theorem*, which they claim proves that if humans have free will, then so too do elementary particles (Conway and Kochen 2006). More precisely, if humans carrying out experiments on elementary particles truly have the ability to make free decisions about what to measure, then the behavior of elementary particles, in response to those measurements, cannot have been purely determined by events in their past history.

At the time, I covered Conway and Kochen's work for *New Scientist* magazine (Merali 2006). My article investigated the theorem's implications for deterministic theories of physics that have been proposed over the years as alternatives to the standard interpretation of quantum mechanics. In particular, I juxtaposed the Free Will Theorem with a deterministic model proposed by physicist Gerard 't Hooft ('t Hooft 2007a). The theorem argues that if we want to retain the notion that humans really are free, we must embrace the indeterminism that lies at the heart of quantum mechanics. Turned on its head, any alternative deterministic theories, such as 't Hooft's, that attempt to do away with the freedom of particles, will inadvertently also rob us of our free will.

The article caused a stir amongst the public and some physicists, leading 't Hooft to publish a new paper responding directly to Conway and Kochen's statements in *New Scientist* ('t Hooft 2007b). In that paper, 't Hooft attempted to redefine "free will" in a way that is compatible with particle determinism. Speaking in a second *New Scientist* article, 't Hooft admitted that the Free Will Theorem had become a stumbling block for his model (Merali 2007). "It's not the mathematics that loses other physicists," he said. "It's the metaphysical worry about free will. Why worry at all about a notion so flimsy as 'free will' in a theory of physics?"

Here I will sketch out what is at stake in the Free Will Theorem and briefly outline the steps of Conway and Kochen's proof of the theorem. I will discuss its implications for physics and for the source of human free will, in light of recent efforts by philosophers and physicists to incorporate quantum mechanical indeterminism into libertarian models of free will (Doyle 2013; Kane 2013). For a full and rigorous mathematical treatment, however, I of course refer you to Conway and Kochen's original writings and their follow-up paper, in which they propose the slightly modified *Strong Free Will Theorem* (Conway and Kochen 2009).

6.2 The Free Will Theorem: What's At Stake?

Quantum mechanics, the theory that governs the behavior of subatomic particles, has become one of the most successful theories in the history of physics, despite predicting a number of paradoxical phenomena. Since its development in the early twentieth century, experiments have repeatedly verified its strangest predictions.

One of the weirdest aspects of quantum mechanics is that, at its core, it is indeterministic: Before measurement, conventional wisdom goes, a quantum particle exists in a superposition of many mutually contradictory states; only after measurement does the particle settle into one of these options. Prior to observation, it is impossible to know with certainty what the outcome of your measurement will be.

Einstein famously rankled at this indeterminism and God's apparent dice-playing with the universe. He believed that particles must contain extra properties—or 'hidden variables'—that do determine their behavior completely. The trouble is that we do not have access to these hidden blueprints; if only we did, we could predict the fate of particles and the outcome of measurements with certainty. Since then there have been many attempts over the years to produce a working deterministic theory (see, for instance, Bohm 1952, and 't Hooft 2007a). Some have been discredited, but others persist, and it is theories such as these that the Free Will Theorem targets—arguing that they are incompatible with human free will.

Conway has stated that neither he nor Kochen originally set out to prove a theorem that undermined hidden-variable theories; they wanted only to understand more about the processes at play in the subatomic world (Conway 2009). Ironically, Kochen had spent many years attempting to develop a deterministic theory in the past, an enterprise that he abandoned upon proving the Free Will Theorem. The theorem cannot prove that deterministic theories are wrong—it may well be the case that such a theory is correct and, if so, we have to accept that free will is an illusion. However, the theorem does make it difficult for those who wish to retain the belief that humans are free agents to support a deterministic theory.

6.3 Proving the Free Will Theorem

The first and most important step of the proof of the Free Will Theorem was actually demonstrated more than 40 years ago by Kochen and fellow mathematician Ernst Specker. It is based on the ways that you can experimentally poke a particle to determine the values of its *spin*—an internal quantum property of the particle that points along a direction. Spin is measured along three perpendicular axes that pass through the center of the particle. Conway likens measuring the components of a particle's spin along three axes to playing the game "20 questions," where one player tries to guess what their opponent is thinking of by asking them yes/no questions. If you play the game honestly, you think of an object before the game begins and stick with it. This corresponds to a deterministic theory—a particle has definite predetermined spin component values assigned along every possible axis before you look at it, and when you probe it along any three axes, it simply yields these set values.

Kochen and Specker tested whether it is possible to set predetermined spin values for a particle along just 33 directions (playing a game of "33 questions" with the particle) before the experiment begins. The catch is, according to quantum mechanics,

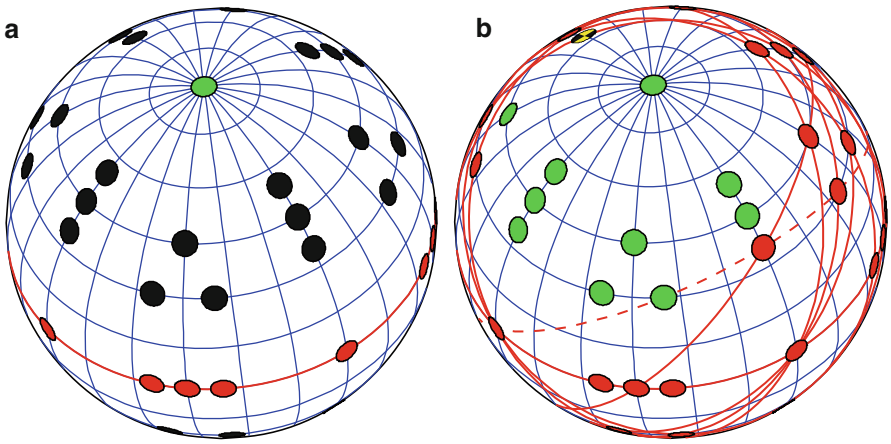


Fig. 6.1 The Kochen–Specker paradox. *Green* dots indicate that the squared spin along the axis passing through the dot and the center of the particle is 0, *red* dots indicate that the squared spin is 1, and *black* dots indicate that the squared spin has yet to be assigned. **(a)** Choose any one of the 33 axes to be *green* (squared spin value 0) as a starting point. By the SPIN axiom, all perpendicular axes are *red* (squared spin value 1). **(b)** Continue to assign spin values according to the SPIN axiom (see appendix). You will eventually reach a point where an axis must take on a value that is simultaneously 0 and 1 (colored *yellow* and *black* here) creating a paradox. (Images courtesy of Jan-Åke Larsson, Linköping University, Sweden (Larsson 2010))

these spin values cannot be preassigned in just any way. There is a constraint, predicted by quantum theory and experimentally verifiable, about the values that measurements along any three perpendicular axes should yield. Kochen and Specker proved that with this constraint in place, you cannot consistently preassign spin values for your particle along 33 directions (Kochen and Specker 1967).

The crux of the derivation of the *Kochen-Specker paradox*, as it is known, is a constraint on the so-called spin-1 particles. If you measure the squared value of the spin of such a particle along three perpendicular axes, you will always uncover the same three values—1,0,1—in various orders. This constraint is the first formal axiom needed to prove the Free Will Theorem:

Spin Axiom: If you measure the squared component of the spin of a spin-1 particle along three perpendicular directions, you will always find the same three values 1, 0, and 1, in some permutation.

Figure 6.1 shows how any attempt to preassign a fixed squared spin value (0 or 1) along these 33 axes will become unstuck, if you are forced to respect the SPIN axiom. (I have included a full walk-through of the derivation of the Kochen–Specker paradox, in which spin values are assigned along each axis, step by step with diagrams, as an appendix.) It turns out that with this restriction in place, you can only set consistent spin values for 30 directions out of a possible 33. The final three directions must take a value that is paradoxically both 1 and 0.

The Kochen–Specker paradox, then, proves that you cannot set values for the squared spin along every possible axis of the particle—you cannot even set them for

33 directions without reaching a contradiction. This result is not a problem for the standard interpretation of quantum mechanics, which embraces indeterminism. In that picture, a quantum particle sets its spin on the fly. This is the equivalent of cheating at "20 questions" (or "33 questions"), where rather than having a set object in mind before the game begins and sticking with it, you keep changing the object in mind, as the questions are being asked.

This result is a big problem for simple deterministic theories, however. There seems to be only one way to avoid hitting a paradox: Nature must somehow restrict the experimenter's choice so that she only chooses to measure along three axes chosen from the 30 directions along which the spin is completely predefined. Nature steers her away from measuring the spin along one of the remaining three paradoxical axes, where the squared spin takes the values of 0 and 1 simultaneously. If that is the case, however, then the experimenter did not truly have free will; her choice of experimental set up was not under her control, but rather was forced by nature to stop her from running into the paradox.

The Kochen–Specker paradox strongly supports the quantum-mechanical indeterministic view and goes a long way towards proving the Free Will Theorem. However, alone it is not enough to rule out a more complex deterministic picture. For instance, the order that the axes are measured could have an effect on the spin values observed. In the 20 questions analogy, although you may not choose one single set object before the game begins, you may still have clearly defined and determined instructions about how to answer the questions before the game starts, depending on their order. For instance, you may decide to answer "yes" to the question, "Is it an animal?" if that is the first question asked, and then "no" to the question, "Is it bigger than a bread basket?" if that is the second question asked, and then to continue to play the game with a mouse as the object in mind. If the questions are asked in the opposite order, however, you may have already decided before the game began to answer "no" to the animal question, then "yes" to the bread-basket question, and to continue playing with a car in mind.

The next steps of the proof of the Free Will Theorem are designed to show that the particle could not return a set of spin values that were predetermined according to a more complicated set of instructions that includes the order that the axes are measured or anything else in the environment and deterministic history of the particle or experimenter.

To prove that environmental factors do not form part of a sophisticated instruction booklet that determines how the particle should behave, requires two more axioms: TWIN and FIN. The TWIN axiom is based on the now famous and experimentally verified quantum phenomenon of "entanglement." Entangled particles are inextricably intertwined, such that a measurement performed on one influences the outcome of the same measurement on the others, no matter how far the particles are separated. More formally:

Twin Axiom: For twinned spin-1 particles, A and B, if an experimenter Alice measures the squared spin for A along three perpendicular axes, x, y, and z, then a second experimenter Bob will measure the same spin values if he performs a spin measurement on B, along the same axes.

The FIN axiom is motivated by the special theory of relativity, which states that information cannot be transmitted faster than the speed of light.[1] Formally:

Fin Axiom: There is a finite upper bound to the speed at which information can be transmitted.

The TWIN axiom allows us to imagine twinned Kochen–Specker-type experiments, carried out by two different experimenters, Alice and Bob, in different laboratories. We'll say that Alice carries out her test on Earth, while Bob is banished to Mars and must perform his test there. Alice, on Earth with her particle A, must pick out three perpendicular directions (out of 33 directions) along which to measure the squared spin.[2] On Mars, Bob will measure the squared spin value along one axis, w, of his particle B.

We want to prove that nothing in particle B's past history can provide information that will allow all possible outcomes of Bob's experiment to be consistently predetermined. So we start by assuming the opposite—that any piece of information in its history *could* set the outcome: Alice's choice of axes to measure x, y, z; Bob's choice of axis, w; or indeed any information in Bob (and B's) past history, labeled β. The rest of the proof involves explaining why these extra factors x, y, z, w, and β cannot help predetermine the outcome of the test.

Alice and Bob perform their tests at roughly the same time. Broadly speaking, we can say that Alice and Bob's laboratories are so distantly separated that, by the FIN axiom, Alice's choice of axes and the order in which she makes her measurements should not influence Bob's result. In fact, Alice may not necessarily have even made her measurements before Bob makes his pick, so that information should not be available to Bob and B. Thanks to special relativity, that constraint can be made even stronger: According to relativity theory, two observers moving relative to each other may not agree on the order in which Alice and Bob carry out their tests. Depending on their motion, it is possible that for one observer, Alice carries out her test first, while according to the other, Bob performs his test first. Given that, it makes little sense to say that the outcome of Bob's test could depend on Alice's choice of x, y, z, or the order in which she makes her measurements.

That still leaves us with the possibility that Bob's result could be fixed by his choice of axis, w, and by other varying information in particle B's past history, labeled β. To eliminate the influence of a varying past history, we can zoom in to the instant just before Bob makes his choice about which direction w to choose. The trick is to take a snapshot of the past history of the universe at that point, collected

[1] Following criticisms to the first version of the Free Will Theorem, the authors replaced the FIN axiom with the MIN axiom, which is experimentally motivated (Conway and Kochen 2009). For our purposes, MIN and FIN perform the same function.

[2] Although the Kochen–Specker setup has 33 directions along which to measure the squared spin, from these there are 40 different "triples"—sets of three perpendicular axes—that Alice could pick. So she is really choosing one triple out of 40 possible triples.

up in "β_0," which is a constant and can thus be removed.[3] That means the only changing parameter left to predetermine the outcome of Bob's spin observation, which we'll call b_0, is the choice of axis, w.

By similar arguments, Alice's results can only depend on x, y, z. We have not yet ruled out that each of Alice's three measurements (one along each axis), which we'll call a_1, a_2, and a_3, could each depend on the order of measurements in some clever and cunning way. But by the TWIN axiom, we can say that if Bob's axis w happens to coincide with Alice's choice x, then a_1 will have the same value as the value of b_0 observed along the x-axis; if w coincides with Alice's choice y, then a_2 will be the same as b_0 observed along the y-axis; and if w coincides with Alice's choice z, then a_3 will equal the value of b_0 along the z-axis.

By the SPIN axiom, we also know that a_1, a_2, and a_3 must be 1, 0, 1 in some order. By TWIN, as just described, we can also then identify the measured values of b_0 along the three axes x, y, and z as 1, 0, 1 in some order. So far, we have seen that b_0 can only be predetermined according to the choice of axis that Bob measures, w. But thanks to TWIN, we have now set b_0 up in a way that forces it to obey SPIN. Our earlier derivation of the Kochen–Specker paradox showed that it is impossible to preset all of b_0's spin values in a way that satisfies SPIN, so once again we have reached a contradiction.[4]

By the Kochen–Specker paradox, then, if Bob's choice was truly free, then the outcome of his measurements was not predetermined. This, finally, brings us to the statement of Conway and Kochen's result:

The Free Will Theorem: *If an experimenter is truly free to choose the directions along which to make measurements of a particle, then the particle's responses must also be free.*

6.4 Criticisms of the Free Will Theorem

The Free Will Theorem, then, proves that if experimenters have free will, so too do elementary particles. Conversely, if the behavior of particles is entirely determined by past events, then we have to kiss goodbye to the notion that humans are free

[3] There are some subtleties that I have glossed over here. It could be argued that the past history of the universe, β, should not simply be dismissed as a constant. It could be that its influence is more complicated, perhaps determined in part by the choices of axes made by Alice and Bob. Conway and Kochen argue that even in this case, it is still possible to rewrite them as constants (Conway 2009).

[4] In the discussion earlier, I have ignored the time (about 0.1 seconds) that it takes for Bob to make his free-will decision. During the time between Bob making his choice and actually hitting the button on his measuring apparatus, some new information from elsewhere in the universe could sneak into his Mars laboratory and influence his results. To this suggestion, Conway and Kochen counter that if this new information is entirely predetermined by past events in the universe (as recorded in β_0) then it can be dismissed by following the same line of argument that allowed them to disregard β_0. Alternatively, if the new information is not entirely determined by the prior history of the universe, then this new information is "free." This new free information could affect the outcome of Bob's observation in some special way, but if so, it reinforces the conclusion that the measured result is influenced by something not predetermined.

agents. As described earlier, the only way that the outcome of quantum experiments can be squared with a deterministic model is if nature conspires to prevent the experimenter from making the measurements that would reveal a paradox, and hence the experimenter's choices are not free.

Some argue that this conclusion does not tell physicists anything that they did not already know about deterministic theories (Goldstein et al. 2010; Gisin 2010). However, Conway and Kochen claim that their result goes beyond previous arguments given by physicists that quantum mechanics must be indeterministic. They say that the Free Will Theorem does not simply say that a certain physical theory (quantum mechanics) cannot predict what a particle will do—a notion that physicists accept—but that if we agree that humans have free will, then the behavior of elementary particles cannot be fully determined by the past, regardless of the physical theory considered. Thus, they are making a statement about the nature of reality, not just about the limitations of a physical theory. Although the SPIN and TWIN axioms are predicted by quantum mechanics, you do not have to buy into the full machinery of quantum theory to accept them. Instead, you can accept SPIN and TWIN on the basis that they can be confirmed in the laboratory, while remaining agnostic about whether the rest of quantum theory correctly describes reality.

In this volume, physicist Antonio Acín argues that the assumption that human experimenters have free will is implicit in the standard interpretation of experiments that demonstrate entanglement and that reject the possibility that signals can be transmitted faster than some finite speed limit (Acín 2013). At the very least, then, it can be said that Conway and Kochen have made this assumption explicit. In independent work, physicists Jonathan Barrett and Nicolas Gisin have also quantified free will and have demonstrated that the results of certain standard quantum experiments would not hold if even a small amount of the experimenters' free will is sacrificed (Barrett and Gisin 2010).

6.5 Conclusion: Deterministic Machines, Random Machines, or Free Agents?

Given the Free Will Theorem, then, what are the implications for physics, for the origins of our choices, and for moral accountability? Conway and Kochen choose to believe that our actions are free—that given a cup of coffee, we have a genuine choice over whether to drink it or to throw it across the room—and so to them, the Free Will Theorem rules out the possibility of finding a deterministic underpinning to quantum mechanics.

Those who still favor a deterministic theory must find a way to reconcile their conception of physics with the loss of libertarian free will. This in itself is not a new endeavor; historically—in particular, before the discovery of quantum indeterminism—much thought has been devoted to this effort (Kane 2013). However, the

Free Will Theorem has provoked recent novel work on free will. As mentioned earlier, 't Hooft, in direct response to Conway and Kochen's work, has attempted to redefine free will in such a way that it is compatible with a deterministic theory of physics, based on his "unconstrained initial conditions postulate" ('t Hooft 2007b).

The new notion of free will in 't Hooft's framework is still restricted and does not allow people the ability to change their minds on a whim. For instance, it is not possible, when drinking a cup of coffee, to freely choose to throw the cup across the room. "I can't change my mind in an instant about whether to drink the coffee or hurl it across the room. My decision must have roots in brain processes that occurred in the past," he has said (Merali 2007). "What's important is that I have freedom to calculate what happens if I throw my coffee cup. Equally, I have the freedom to calculate the effects after I drink from my cup." But, in this formulation, we necessarily lack the freedom to instantaneously switch between which of these initial states we start from. 't Hooft's reformulation of free will has also been criticized because it requires a well-defined causal relationship at the fundamental level. However, sophisticated quantum experiments suggest that there is no such time ordering at a deep level (Suarez 2007; see also Merali 2007 and Merali 2011 for popular discussions).

Even if we do embrace quantum indeterminism, however, it remains unclear that this is enough to explain the true origins of human free will. By the Free Will Theorem, the behavior of elementary particles may be "free" in some sense, but there is no apparent "will" involved in their choices. Philosophers such as Tim Maudlin have argued that this is not true freedom, only randomness. In this sense, quantum mechanics is no better than a deterministic theory in terms of making us morally responsible for our choices: "For philosophers, both arguments can be troubling. Quantum randomness as the basis for free will doesn't really give us control over our actions," Maudlin has said. "We're either deterministic machines, or we're random machines. That's not much of a choice" (Merali 2006).

Conway has countered that free will rooted in quantum indeterminism should not be identified with randomness. He uses the example of a backgammon tournament, where many competitors arrive at a venue and are divided into pairs to play games in parallel. The opening move for all games is set by a single roll of a pair of dice by the tournament organizer. In this way, all opening players must make their first move based on the random outcome of that dice roll. However, each opening player has the free choice to decide which pieces to move, and in what way, in accordance with that random dice roll. In this sense, randomness leaves room for "freedom of the will."

Other philosophers have also gone some way toward addressing the criticism presented by Maudlin, by arguing that indeterminism can play a role in the mechanism of free will, without rendering all decisions random and, in turn, robbing humans of moral accountability. In this volume, for instance, Robert Doyle outlines two-stage solution to the problem of free will (Doyle 2013). In the first stage of his model, quantum indeterminism provides an initial random seed that generates a set of alternative possibilities to be considered. In the second stage—a stage of

"adequate determinism"—the will acts to decide which of these possibilities to choose, using reasoning that is based on the agent's past experiences and personality.

Also writing in this volume, Robert Kane describes a model in which key "self-forming actions" in a person's history are not fully determined (Kane 2013). These self-forming actions occur when a person is faced with an important decision that causes them to wrestle with conflicting motivations. Indeterminism, in this picture, is generated during these internal conflicts, such that the agent's final decision is unpredictable during deliberation. In order to make a choice, the agent must make an effort to overcome the indeterminism that acts as an obstacle to his choice.

Both Doyle and Kane's models succeed in invoking indeterminism to break the causal chain—that begins with the production of particles during the Big Bang—that would otherwise fully determine the outcome of every choice in the history of the universe. In both cases, this injection of indeterminism escapes Maudlin's criticism that choices become simply random, because willed deliberation is involved in the decision-making process. This is in line with Conway's argument that free will based on indeterminism is not necessarily reduced to randomness.

In fact, Conway and Kochen go further, arguing that the Free Will Theorem proves that the opposite of determinism is not randomness, but a third option: freedom, as *defined* by the behavior of elementary particles. Here "free answers" are characterized by being chosen on the fly; they cannot have been decided upon ahead of time. As such, they stand in opposition to not only predetermined answers, but also to random answers because decisions set by the random roll of a dice could have been set at an earlier time without affecting the observed outcome. In the backgammon example, it makes no difference to the play if the dice were rolled at the beginning of the tournament, a week earlier and recorded, or at the beginning of the universe. In this picture, then, random numbers are no better than predetermined values.[5] Since the "free" behavior of quantum particles, described this way, is neither deterministic nor random, it could provide a mechanism for making free choices for which we can be held morally accountable.

This identification of the essential feature of true freedom with the ability to make a decision on-the-fly resonates with aspects of both Doyle and Kane's models. Both argue that the crucial characteristic of free will is not that decisions are entirely divorced from factors in the past history of the agent—on the contrary this is necessary to ensure that the agent can be held accountable for *willing* her choices—but that the agent's choice is open up to the last moment of her deliberation (Doyle 2013; Kane 2013).

Returning to the question posed in the title of this paper, "Are humans the only free agents in the universe?," the answer, according to the Free Will Theorem, appears to be "no, so too are elementary particles"—but only if we accept the

[5] The assertion that randomness is as bad as determinism when trying to formulate a fundamental physical theory has been attacked. See for instance, Goldstein et al. (2010).

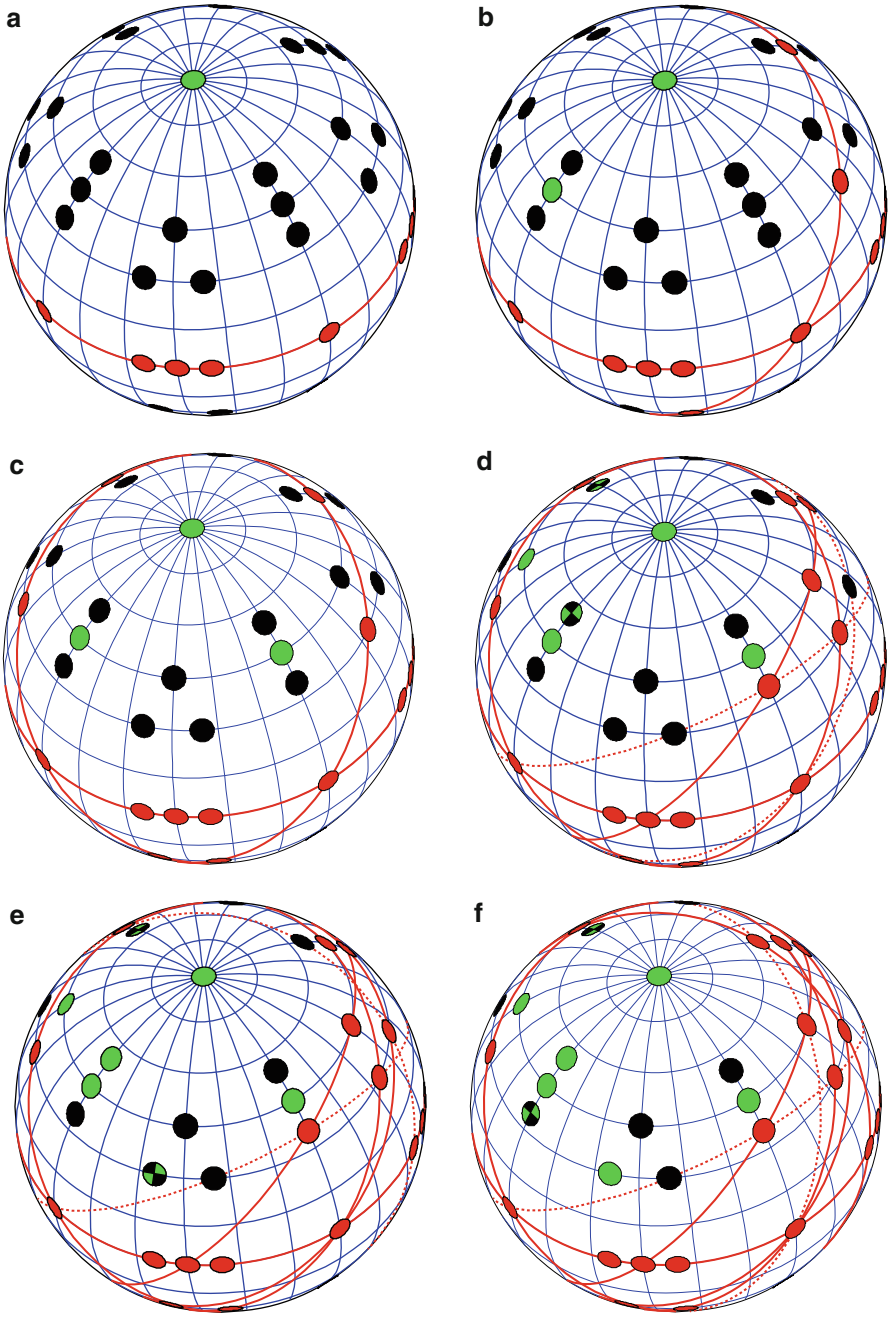


Fig. 6.2 (continued)

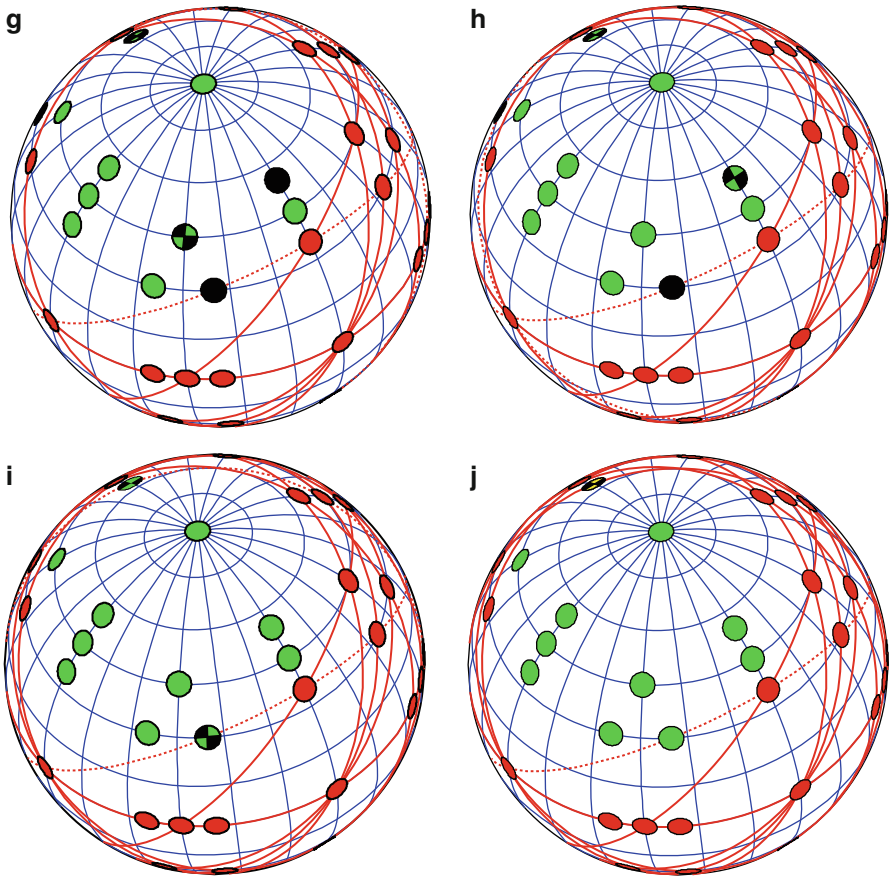


Fig. 6.2 (Visualizing the Kochen–Specker paradox. **(a)** Step 1: Choose any one of the axes to be *green* as a starting point. By Rule 2, all perpendicular axes are *red*. **(b)** Step 2: Choose one of the *red* axes on the equator. Two directions on the 45° latitude are orthogonal to this direction. By Rule 2, one of these must be *green*, so choose one. All perpendicular directions can then be set to be *red*. **(c)** Step 3: Repeat the argument in Step 2, to color another direction *green*, and all corresponding perpendicular directions *red*. **(d)** Step 4: Repeating the argument of Step 2 again, gives another *green* axis, and another set of perpendicular *red* axes. There are now a number of *red* axes running through the particle. Some of these *red* axes are perpendicular to each other, and so, by Rule 3, they force their third perpendicular direction to be *green* (marked as checkered *green*/*black* here for clarity). **(e)** Step 5: Choose one of the checkered dots to be *green*. (Save the other checkered dot for later.) This choice will force another perpendicular direction to become *red* (by Rule 2), which in turn, forces a new *green* direction, by Rule 3 (marked as checkered *green*/*black* here). **(f)** Step 6: Repeat Step 5, coloring this axis *green*, and forcing a new axis to turn *green*/*black*. **(g)** Step 7: Repeat Step 5 again, creating another new *green*/*black* axis. **(h)** Step 8: Repeat Step 5 again, creating another new *green*/*black* axis. **(i)** Step 9: Repeat Step 5 again, creating another new *green*/*black* axis. **(j)** Step 10: Attempt to repeat Step 5 again. This time, however, when you color the next *green*/*black* axis *green*, you will force the remaining *green*/*black* axis to be *red*. So there is a contradiction as at least one axis must be both *red* and *green* (here marked *yellow*/*black*). (Images courtesy of Jan-Åke Larsson, Linköping University, Sweden (Larsson 2010))

assumption that humans themselves are free. Conway and Kochen believe that their theorem motivates a belief in free will, by intimately connecting it with a surprising discovery by physicists (that is, with quantum indeterminism). However, by Conway's own admission, the theorem cannot logically disprove the idea that free will does not exist: "Our lives could be like the second showing of a movie—all actions play out as though they are free, but that freedom is an illusion," he has said (Merali 2006). I cannot promise that humans are free agents either. So I think that the correct answer to the opening question about whether there are other non-human-free agents in the universe is only: "maybe." But then, perhaps I had no choice other than to reach that conclusion.

Acknowledgements This paper was born from two popular articles that I wrote for *New Scientist* magazine in 2006 and 2007, discussing free will and deterministic theories of physics. I must thank the following people for spending many hours describing their own work and the research of others to me, during the development of those articles: John Conway, Simon Kochen, Gerard 't Hooft, Hans Halvorson, Tim Maudlin, and Antoine Suarez, who also gave me the opportunity to speak about this topic at the *Is Science Compatible with Our Desire for Freedom*? meeting in Barcelona, Spain, in October 2010 and invited me to contribute to this volume. I would also like to thank my former editors Anil Ananthaswamy, Rowan Hooper, and Matt Walker at *New Scientist* magazine, who agreed that these issues were interesting and important enough to merit a large amount of coverage in a popular science publication and who helped to frame my magazine articles.

Appendix: The Kochen-Specker Paradox

The Kochen-Specker paradox demonstrates that it is impossible to preassign spin values along every possible direction for a particle (specifically a spin-1 particle), in a way that can satisfy the SPIN axiom. In fact, it is impossible to preassign spin values along just 33 directions in a way that satisfies this rule.

The SPIN axiom states that if you measure the squared component of the spin for such a particle along three perpendicular directions, you will always get the same three values—1, 0, and 1—in some order. This can be visualized by attempting to preassign a fixed spin value (0 or 1) along 33 directions. In Fig. 6.2, green dots indicate that the spin along an axis passing through the dot and the center of the particle is 0, red dots indicate that the spin is 1, and black dots indicate that the spin has yet to be assigned.

The SPIN axiom can be broken down into three simpler rules:

Rule 1: Opposite directions (dots on the particle) always have the same spin value.
Rule 2: Two perpendicular directions cannot both be 0.
Rule 3: Three perpendicular directions cannot all be 1.

Using these rules, we can attempt to set spin values for all directions. However, we will find that some directions must paradoxically be both red and green.

References

Acín, A. (2013). True quantum randomness. In A. Suarez & P. Adams (Eds.), *Is science compatible with free will (chapter 2)*. New York: Springer.

Barrett, J. & Gisin, N. (2010). How much measurement independence is needed to demonstrate non-locality? arXiv:1008.3612v2.

Bohm, D. (1952). A suggested interpretation of the quantum theory in terms of "hidden" variables II. *Physical Review, 85*, 180–193.

Conway, J. H., & Kochen, S. (2006). The free will theorem. *Foundations of Physics, 36*, 1441–1473.

Conway, J. H., & Kochen, S. (2009). The strong free will theorem. *Notices of the American Mathematical Society, 56*, 226–232.

Conway, J. H. (2009). Free will lecture series, Princeton University Web Media—Lectures. Retrieved from http://www.princeton.edu/WebMedia/lectures/ on 28 September 2012

Doyle, R. (2013). The two-stage solution to the problem of free will. In A. Suarez & P. Adams (Eds.), *Is science compatible with free will (chapter 16)*. New York: Springer.

Gisin, N. (2010). The free will theorem, stochastic quantum mechanics and true becoming in relativistic quantum physics. http://arxiv.org/abs/1002.1392

Goldstein, S., Tausk, D. V., Tumulka, R., & Zanghi, N. (2010). What does the free will theorem actually prove? *Notices of the American Mathematical Society, 57*, 1451–1453.

Kane, R. (2013). Can a traditional libertarian on incompatibilist free will be reconciled with modern science? Steps towards a positive answer. In A. Suarez & P. Adams (Eds.), *Is science compatible with free will (chapter 17)*. New York: Springer.

Kanfer, S. (1997). Isaac Singer's promised city. *City Journal*. Retrieved from http://www.city-journal.org/html/7_3_urbanities-isaac.html on 28 September 2012.

Kochen, S., & Specker, E. (1967). The problem of hidden variables in quantum mechanics. *Journal of Mathematics and Mechanics, 17*, 59–87.

Larsson, J.-A. (2010). Visualizing the Kochen-Specker paradox. Retrieved from http://people.isy.liu.se/jalar/kochen-specker/ on 28 September 2012.

Lucretius. (1996). *De Rerum Natura*, Translated by W. H. D. Rouse. Cambridge, Massachusetts: Harvard University Press. 2.250–256.

Merali, Z. (2006, May 6). Free Will—You only think you have it. *New Scientist magazine*, 8–9. Retrieved from http://www.newscientist.com/article/mg19025504.000-free-will–you-only-think-you-have-it.html

Merali, Z. (2007, August 4). Entangled in the free will debate. *New Scientist magazine*, 10–11. Retrieved from http://www.newscientist.com/article/mg19526154.200-free-will–is-our-understanding-wrong.html on 28 September 2012.

Merali, Z. (2011, March). Physics of the Divine. *Discover magazine*, 48–53. Retrieved from http://discovermagazine.com/2011/mar/14-priest-physicist-would-marry-science-religion on 28 September 2012.

Suarez, A. (2007). Classical demons and quantum angels: On 't Hooft's deterministic quantum mechanics. http://arxiv.org/abs/0705.3974

't Hooft, G. (2007a). A mathematical theory for deterministic quantum mechanics. *Journal of Physics: Conference Series*, *67*, 012015.

't Hooft, G. (2007b). The free will postulate in quantum mechanics. arXiv:quant-ph/0701097v1.

Chapter 7
The Origin of Freedom in Animal Behaviour

Martin Heisenberg

Abstract Behaviour of humans and most animals can be free. Behaviour is free if the subject does of her/his own accord what must be done. Action selection is the main task of a brain. The search for the right behaviour is demanding because it requires assessing the possible consequences of the available behavioural options. Freedom is reduced externally, if fewer adaptive behavioural options are available or internally, if the search process in the brain is impaired. Most animal societies enforce cooperation at the expense of individual freedom. In contrast, human societies can base cooperation on shared intentions. In this way the individuals can cooperate without sacrificing behavioural freedom.

Keywords Self • Autonomy • Chance • Initiating activity • Open future • Shared intentions

7.1 Introduction

Recent claims that freedom is an illusion or self-deceit have attracted much attention (e.g. Wegner 2002; Roth 2004; Singer 2006). Some scientists and philosophers maintain that a behaviour causally determined by natural law cannot be considered free, and a behaviour that is released by chance is not free either. Lawfulness and chance, they say, are an alternative that leaves no room for anything else. Therefore, if neither lawfulness nor chance allow for freedom, how could it possibly be real. This seemingly "waterproof" argument is flawed. Chance and lawfulness occur together and even depend upon each other. Their specific interplay constitutes our world in which the future is open and creation has not ceased. It is in this world, where behavioural freedom has evolved.

M. Heisenberg (✉)
Rudolf-Virchow-Centre, University of Würzburg, 97080 Würzburg, Germany
e-mail: Heisenberg@biozentrum.uni-wuerzburg.de

A. Suarez and P. Adams (eds.), *Is Science Compatible with Free Will?: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience*,
DOI 10.1007/978-1-4614-5212-6_7,

7.2 Behavioural Freedom

Freedom comes in many forms and under many circumstances. We can be free *for* -, and free *from* - something. There are free decisions, free will, creativity, freedom of thought, faith, speech, free commerce, a free press and many other kinds of freedom. They all relate in one way or another to behaviour. So, let us consider freedom in behaviour. We all know what that is. We can do this or that. We have behavioural options with sufficiently positive prospects and can make use of them or not. A government banning travel abroad or preventing access to certain domains of the Internet diminishes the behavioural freedom of its citizens. Of behavioural freedom one may have too much and too little. It is vividly felt, if denied. If for any reasons behaviour could, by first principles, never be free, none of the other kinds of freedom would likely be real. Behavioural freedom is a freedom *for* something: for living your own life.

Freedom has come about like any other property of living organisms in the course of evolution. It is a quality of behaviour. It is so old and so basic that we share it with most animals. We can let an animal free that had been trapped, and this is not a metaphorical expression. The animal had been deprived of some of its behavioural options. In describing how behavioural freedom is possible, I will draw some of my examples from animals.

7.3 Behavioural freedom and consciousness

What is this special property that allows behaviour to be free? It surely is not consciousness. For my actions to be free, I do not have to be conscious of them. To use a dramatic example: If I am driven into a corner of a dark city by a gangster I panic because I feel the rapid loss of freedom, the diminishing number of behavioural options with potentially beneficial outcomes, but this happens independent of whether I reflect about the situation or not. Looking at the problem from the other side, if I have a neurosis that forces me to touch the two sides of each door frame in passing through, I am unfree in this regard, irrespective of whether I am aware of performing this act or not. Once again, whether a behaviour is free or unfree does not depend upon me being conscious of it. Conscious reflection may improve a difficult decision, as may meditation or discussion, but it does not even necessarily do so.

7.4 The initiating quality of behavioural activity

Behavioural freedom must have evolved with brain and behaviour. To see this one has to understand how the brain organizes behaviour. Behaviours can be actions or reactions. We may disregard the latter. Reactions may well be unfree. They are triggered by stimuli from outside and often occur in emergencies. Let us consider

the actions. Most behaviours are actions. Behaviour is active. It originates in the animal or human displaying it. Typically its generation involves elements of chance in the timing and selection process. Without any chance in the selection process behaviours could always be traced back to sufficient causes from outside the organism. A typical example of an action is trying out. Actions are more likely to occur when the animal is relaxed. I have called this activity "initiating activity" (Heisenberg 1983, 1994) to highlight what distinguishes it from re-activity and rhythmic activity.

The initiating quality of behavioural activity is intuitively obvious. It is more apparent in insects than in mammals. The active voice in our language refers to it. The little girl jumps down from the chair. She has the impulse to do so by the active nature of her behaviour. We do not say: "Some hidden stimuli or her nerves and muscles made her jump". She jumps of her own accord. The jump has the quality of a beginning. We all see ourselves and others as agents in our respective affairs, as the originators or authors of our behaviour and its consequences. Moreover, we introspectively experience initiating activity in our thinking.

Active behaviour has long been ignored or even denied among behavioural scientists. In order to live up to the "exact sciences" and with the maxim "Nothing comes from nothing!" it was explained away by assuming a lack of knowledge on the part of the observer about the stimuli. Sensory-motor reflexes such as the eye blink or the Patella reflex were considered the basic building blocks of behaviour (Sherrington 1919). This view had been greatly popularized by the discovery of the so-called sign stimuli, sensory signals that do, indeed, elicit complex behaviours (e.g. Lorenz 1965). Even today it is still occasionally proclaimed that the principal task in brain science is to understand how the multitude of sensory stimuli is transformed into motor commands. Meanwhile, though, roles of chance in brain and behaviour are beginning to be recognized (Glimcher 2005; Herz 2007; Maye et al. 2007; Vaziri and Plenio 2010).

Presumably, the oldest form of behaviour is self-motion, the active change of position in space. It goes back to the prokaryotes and seems to have been reinvented several times independently with the advent of multicellular organisms. The stochastic element in the initiation of locomotion is still apparent in the random walk of bacteria (Adler 1975), the head turning sequence of crawling *Drosophila* larvae (Gomez-Marin et al. 2011), random search (Viswanathan 1996, 1999) and predator–prey relationships (discussed in Maye et al. 2007). Activity in higher animals has largely receded to the brain and we observe it as an indispensable element of behavioural organization, as will be discussed below.

7.5 Initiating activity and the "self"

In order to understand initiating activity we have to introduce the concept of "self". The biosphere is subdivided into organisms. An organism is a highly autonomous system, a small cosmos separated from the rest of the universe. The term "self"

assigns something to this organism in the perspective of this organism. We meet the distinction between self and non-self at many levels; take restriction enzymes, RNAi, mirror neurons, graft rejection, sociobiology or psychology. In common everyday language we attribute "self" only to subjects, i.e. humans and animals.

What matters in our case is that the organism generates the behaviour her- or him*self*. Active behaviour is not released by external stimuli. There are no sufficient causes outside the organism to make the organism release the particular behaviour. The organism initiates it from within. This is how actions are defined. Why is it important that a behaviour is initiated by the organism displaying it? For the same reason that Darwin superseded Lamarck. Most situations are partially new. In the sensory-motor reflex the organism is already prepared to meet a potential challenge, whereas behavioural activity deals with the unforeseen, with challenges for which the organism has no ready-made answers yet. Finding the right behavioural option is a demanding task and sometimes even a creative process.

7.6 Research on flies

In this and the following four paragraphs I will describe research on flies (*Eristalis*, *Drosophila*) that led to the concepts of initiating activity and behavioural freedom advocated here. Characteristic of behavioural activity is that the organism normally does not respond to the sensory stimuli it causes by its own behaviour. A well-known example from the human sphere is the fact that one normally does not notice the floating or jumping visual surround during one's eye movements or that one cannot tickle oneself. The profound difference in processing a sensory stimulus depending upon whether it is *self*-induced or externally generated has been formalized as the "Principle of Reafference" by E. von Holst and H. Mittelstaedt (1950). They observed hoverflies walking in an arena surrounded by a cylindrical wall of black and white vertical stripes. As long as the cylinder was standing still, the animals seemed entirely unrestrained by their visual surround in turning left and right, but if the cylinder was set in motion (as if the animals were—miraculously—turned by an external force despite their tarsal contact with the ground) they tried to stabilize their orientation in space by turning in register with the moving stripes. To exclude the possibility that the flies just blocked the entire visual input during locomotion, the experimenters rotated the heads of the animals by 180 deg. such that the left eye was at the place of the right one and the proboscis pointed upward. As soon as these animals started walking in the stationary drum, they got into violent uncontrollable pirouettes showing that they did perceive the relative motion of the surround. The authors concluded that self-generated turning comes with the expectation of a visual

motion signal in the opposite direction to that of the turn and that the flies perceive this motion signal but normally do not respond to it. Externally imposed turning, on the other hand, is likely to require compensatory action. While the Principle of Reafference was quickly accommodated by control-theoretical approaches, its radical departure from the stimulus–response concept was largely ignored.

We followed up on these experiments using a flight simulator in which the tethered fly (*Drosophila*) could control the angular velocity of the panorama by its yaw torque. Once we got the parameters of the artificial feedback loop adjusted to the strength and dynamics of *Drosophila* flight, we were struck to find that flies could instantly (within 30 msec) distinguish whether a motion stimulus was self-generated or externally applied. Like hoverflies, *Drosophila* ignored any amount of visual motion that had the expected direction against its own intended turns, but violently reacted to the motion signals if we artificially inverted their direction. This shows that the distinction between self and non-self in lower animals is a robust phenomenon that shows even under observationally sharpened, highly reduced laboratory conditions (Heisenberg and Wolf 1984).

To mention another even more telling example, one can let the tethered fly control the ambient temperature with its yaw torque—a situation never before experienced by this fly or its ancestors. For instance, yaw torque to the left causes a pleasant 25°C, while yaw torque to the right instantaneously raises the temperature to a dangerous 42°C. At the start the fly cannot know that it is its own yaw torque that causes the switch. To find out, the fly has to activate the behavioural modules it has available in this restrained situation and has to register whether one of them might have an influence on the temperature. The fly cannot wait for an appropriate sensory stimulus from outside to elicit the respective behaviour. It must have a way to trigger its behaviours it*self*, in order to correlate these events with the changes in temperature (Heisenberg et al. 2001). The fly brain is built such that under certain circumstances the items of the behavioural repertoire can get released independent of sensory stimuli.

Over the last 30 years we have found many behavioural manifestations of activity in the fly brain. Like heat, flies can control also odour intensity with their yaw torque. They can control the angular velocity of a panorama surrounding them not only by yaw torque but also by forward thrust, body posture or abdomen bending. In ambiguous sensory situations they actively switch between different perceptual hypotheses, they modify their expectations about the consequences of their actions by learning, and they can actively shift their focus of attention restricting their behavioural responses to parts of the visual field. In all these behavioural tasks the fly is trying out. The respective behaviour is generated because of its potential consequences for the organism. The mechanism initiating it must be under the control of brain centres evaluating the consequences of the behaviour (Wolf and Heisenberg 1991).

7.7 Animals and humans generate behaviour by themselves

Initiating activity serves situations in which the fly does not know yet what to do. Little is known about how the search process is organized. There are many ways in which chance can have an adaptive role in it, besides structuring the temporal sequence of activations. For instance, the fly may invent a new behavioural option. Several studies in flies indicate that stochastic processes are indeed involved (Brown and Haglund 1994; Martin et al. 2001; Maye et al. 2007; Gomez-Marin et al. 2011). Not all chance events in the brain must result immediately in behaviour. Some may be eliminated by deterministic "selection" processes before their execution. What matters is that the fly cannot know the solutions to most real-life problems. Its repertoire of behavioural modules is all it has available to find out.

In order to account for initiating activity, one has to acknowledge the "self". Animals and humans generate behaviour by them*selves*. The concept of "self" is the main reason to insist on objective chance. It could be said to make no sense to assign a behaviour to an organism if, on basic principles, any behavioural activity could be traced back through a nearly infinite chain of causations to the beginning of the Universe. An animal or human being is the author of a behaviour, as long as no sufficient causes for this activity to occur are coming from outside the organism. Authorship is crucial for behaviour. Behaviour can have good or bad consequences. It is the author for whom the consequences matter the most and who can be held responsible for them.

7.8 Freedom and authorship

This finally brings us back to behavioural freedom. What is at stake with freedom is the quality of behaviour. My behaviour is free, if it is indeed my own and if it is adaptive. Or, following I. Kant (1783), free is who does of his own accord what has to be done. Initiating activity accounts for "*. . .of his own accord. . .*", for authorship. How could my behaviour serve me well, if it were not my own? With the second part, the "*. . .what has to be done*" Kant explicitly invokes the high quality of behaviour for it to be called free. Random behaviour is not free. As a philosopher Kant refers to the moral law but for present-day biologists "adaptive behaviour" may do. Free is who does of her/his own accord what is adaptive. To do what has to be done requires an intact brain and motor system, a thorough choice process and the right opportunities.

In the discussion of freedom we are missing yet another property of brain and behaviour: its uniqueness. While it is a cheap truism that everything (except atoms and molecules) is unique, the uniqueness of organisms becomes more and more significant with their increasing complexity and autonomy. The uniqueness of the human individual is highly valued, as love poems and the threat of cloning humans testify.

There are neither two brains nor two life histories that are identical. Behaviour is a highly personal affair. Given the inter-individual differences, only the subject itself can find in a particular situation the one behavioural module that is the most appropriate for her/his goals.

7.9 Freedom and social context

If freedom is a natural property of brain and behaviour, why, then, has this topic come up so late in evolution and only in human affairs? Why is freedom a centrepiece of Christian faith and something like the holy grail of western civilization? For an answer we have to turn to sociality. Among humans the issue of freedom occurs predominantly in the social context. This should be no surprise. If the social group is a "super-organism" its members lose some of their autonomy. Indeed, in animal societies from ants to apes this is what one observes. The needs of the group are imposed upon the individual, at the expense of individual behavioural freedom.

Homo sapiens probably is the most social of all species. We owe this superlative to a unique development early in hominid evolution. In human societies communication and cooperation can rely on shared intentionality, on the common goals, preconceptions and values of their members (Tomasello 2008; Hamann et al. 2011). The quest for freedom deals with the conflict between the atavistic kind of sociality inherited from our animal ancestors and the more recently evolved specifically human kind. It reminds us that even in cooperation the quality of our behaviour deteriorates if we are forced rather than convinced or persuaded. Behaviour forced upon us is not our own.

This consideration also shows why in species such as *Drosophila* freedom is not as important an issue as in humans: The quality of fly behaviour is not compromised as much by the fly's social interactions as is that of human behaviour. Flies mostly can do of their own accord what has to be done.

7.10 Could a robot be free?

Could we build freedom into the "behaviour" of a robot? Why not? We might be tempted to argue that a limited amount of it has been implemented in the Mars robot which cannot be tightly controlled from earth because of the time delay (Matijevic 1998). In unforeseen situations the robot has to evaluate the potential consequences of its behavioural options to activate the right one. But, perhaps this quick answer is naïve. To which extent has the robot a self? How substantially different is the autonomy of a robot from that of an organism? Could one build a robot that would not, in one way or another, reflect the goals of the engineer who built it, even if some chance processes had been implemented in its operating electronics?

7.11 Conclusion

Behavioural freedom is natural. It is an element of human and animal behaviour. We can understand how it is possible at all. It depends upon the interplay of chance and lawfulness. Due to this interplay a behaviour can originate in an organism, the subject. The subject is the author of her/his behaviour. Aims, motivation, intentionality, creativity and trying-out are all built upon this authorship. Kant's (1783) definition of freedom in terms of "doing, of one's own accord, what must be done", shows that the real topic of freedom is not physics or neurobiology but the self, the subject in its autonomy and integrity. Our freedom is not obviated by natural law. It is threatened by the atavisms in our social relations. Others still impose their will upon our behaviour, while the specifically human way of cooperation is based on shared intentions. This is why we strive for freedom.

References

Adler, J. (1975). Chemotaxis in Bacteria. *Annual Review of Biochemistry, 44*, 341–356.

Brown, W., & Haglund, K. (1994). The Landmark Interviews. *Bringing behavioural genes to light. Journal of NIH Research, 6*, 66–73.

Glimcher, P. W. (2005). Indeterminacy in Brain and Behaviour. *Annual Review of Psychology, 56*, 25–56.

Gomez-Marin, A., Stephens, G. J., & Louis, M. (2011). Active sampling and decision making in Drosophila chemotaxis. *Nature Communications*, *2*, 441 | DOI: 10.1038/ncomms 1455.

Hamann, K., Warneken, F., Greenberg, J., & Tomasello, M. (2011). Collaboration encourages equal sharing in children but not chimpanzees. *Nature, 476*, 328–331.

Heisenberg, M. (1983). Initiale Aktivität und Willkürverhalten bei Tieren. *Naturwissenschaften, 70*, 70–78.

Heisenberg, M. (1994). Voluntariness (Willkürfähigkeit) and the general organization of behaviour. In: *Flexibility and Constraint in Behavioral Systems, (R.J. Greenspan & C.P. Kyriacou, eds.),* pp147-156, Wiley & Sons Ltd.

Heisenberg, M. & Wolf, R. (1984). *Vision in Drosophila. Studies of Brain Function Vol. XII,* V. Braitenberg, Ed., Springer Berlin, Heidelberg, New York.

Heisenberg, M., Wolf, R., & Brembs, B. (2001). Flexibility in a single behavioural variable of Drosophila. *Learning & Memory, 8*, 1–10.

Herz, A.V.M. (2007). Neuronaler Determinismus: Nur eine Illusion? In: *Naturgeschichte der Freiheit (ed. J.-C. Heilinger),* pp35-42, Berlin New York, Walter de Gruyter.

von Holst, E., & Mittelstaedt, H. (1950). Das Reafferenzprinzip. *Wechselwirkungen zwischen Zentralnervensystem und Peripherie. Naturwissenschaften, 37*, 464–476.

Kant, I. (1783). Prolegomena zu einer jeden künftigen Metaphysik, die als Wissenschaft wird auftreten können. *Akademie-Ausgabe Bd. IV*, p338, Berlin 1910.

Lorenz, K. (1965). *Über tierisches und menschliches Verhalten, Vol. 2* (e.g. pp.210-211) Munich: Piper.

Martin, J. R., Faure, P., & Ernst, R. (2001). The power law distribution for walking-time intervals correlates with the ellipsoid body in Drosophila. *Journal of Neurogenetics, 15*, 205–219.

Matijevic, J. (1998). The pathfinder mission of Mars: Autonomous Navigation and the Sojourner Microrover. *Science, 280*, 454–455.

Maye, A., Hsieh, C. H., Sugihara, G., & Brembs, B. (2007). *Order in spontaneous behaviour. PLoS ONE, 2*, e443.
Roth, G. (2004). Das Problem der Willensfreiheit aus Sicht der Hirnforschung. In: *Debatte, Heft 1*, pp83-92, Berlin-Brandenburgische Akademie der Wissenschaften.
Sherrington, C. S. (1919). *Mammalian physiology*. Oxford: Clarendon Press.
Singer, W. (2006). Neurobiologische Anmerkungen zum Freiheitsdiskurs. In: *Debatte, Heft 3*, pp17-26, Berlin-Brandenburgische Akademie der Wissenschaften.
Tomasello, M. (2008). Origins of Human Communication. MIT Press.
Viswanathan, G. M., Afanasyev, V., Buldyrev, S. V., Murphy, E. J., Prince, P. A., & Stanley, H. E. (1996). Levy flight search patterns of wandering albatrosses. *Nature, 381*, 413–415.
Viswanathan, G. M., Buldyrev, S. V., Havlin, S., da Luz, M. G. E., Raposo, E. P., & Stanley, H. E. (1999). Optimizing the success of random searches. *Nature, 401*, 911–914.
Wegner, D.M. (2002). The illusion of conscious will. MIT press.
Wolf, R., & Heisenberg, M. (1991). Basic organization of operant behaviour as revealed in Drosophila flight orientation. *Journal of Comparative Physiology (A), 169*, 699–705.

Part II
Neuroscience and Free Will

Chapter 8
The Role of Inhibitory Control of Reflex Mechanisms in Voluntary Behavior

Flavio Keller and Jana M. Iverson

Abstract Gaze is often a powerful cue as to where someone's attention is directed and as to what someone intends to do. However, the relationship between fixational eye movements, attention, and intentions is not always straightforward. The phenomenon of covert attention, by which we can direct attention to visual objects that are not being foveated, demonstrates that visual attention can be uncoupled from eye fixations. Observations such as these suggest that eye movements are an example of interaction between reflexive and voluntary behavior. Shifts of selective visual attention are controlled in part by the same frontal areas that control voluntary eye movements. The role of voluntary inhibition of reflex eye movements is clearly shown in the antisaccade task, in which participants learn to look away from a salient stimulus that would trigger a reflex saccade. Voluntary inhibition of reflex behavior in humans appears to be a prerequisite for the emergence of free will.

Keywords Gaze • Saccade • Inhibition • Visual attention • Free will

8.1 Introduction: The role of gaze in judgments about voluntary behavior

If we see a man struck by a falling vase of flowers while walking under a window, we may look up at the window to see whether someone pushed the vase. If we were to see a person at the window, we would start wondering whether he had intended to hit the man passing below by voluntarily pushing the vase. For this judgment about

F. Keller (✉)
Lab of Developmental Neuroscience, Università Campus Bio-Medico, Rome, Italy
e-mail: f.keller@unicampus.it

J.M. Iverson
Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA
e-mail: jiverson@pitt.edu

A. Suarez and P. Adams (eds.), *Is Science Compatible with Free Will?: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience*,
DOI 10.1007/978-1-4614-5212-6_8,

intention and wilful behavior, gaze is a crucial element: if we happened to see that he threw the vase aiming at the man passing by (as evidenced by coordination between arm movement and gaze), we would have no doubt that he wanted to hit the man. But if we had seen the same man performing exactly the same arm movement while looking elsewhere, we would be inclined to think that he did not want to hit anyone.

This example illustrates the fact that gaze is an important cue for our judgments about intentions and voluntariness of actions. There is experimental evidence indicating that even very young children attend to gaze to make connections between people and the objects of their gaze. For instance, by 12 months infants appear to utilize gaze to encode relationships between actors and objects and are sensitive to changes in these relations; they look longer at a toy that has been the target of an actor's gaze, even when that toy is no longer the focus of the actor's gaze (Woodward, 2003).

Why is gaze such a powerful cue for intentions? We will argue that this is the case because inhibiting reflex responses is particularly difficult in the oculomotor system. We begin by briefly reviewing some fundamental principles of oculomotor control and neural mechanisms that subserve voluntary versus reflexive eye movements. Next, we describe top-down and bottom-up mechanisms that guide ocular saccades. We then discuss the relationship between saccades and visual attention; and in a final section, we examine some inhibitory mechanisms that permit us to shift fixation away from visual objects that are powerful stimuli for reflex saccades.

8.2 Characteristics of eye movements. Bottom-up and top-down control

In contrast to all other senses (hearing, touch, etc.), the sense of sight is characterized by simultaneity of presentation. The sense of sight has been traditionally associated with theory (*vita contemplativa*) in contrast to praxis (*vita activa*; Jonas, 2001). However, research over the past 20 years has revealed a unique and intimate relationship between the sense of sight and eye movements. No other organ of sense shows such tight links between perception and action. In fact, of all types of bodily movements, eye movements are fully dedicated to one purpose (sight), while limb movements can be dedicated to different purposes (e.g., touch, manipulation of objects and instruments, locomotion). It has been calculated that eye movements are the most frequent movements that humans make; indeed, they are more frequent than heartbeats (Ingram and Wolpert, 2011).

From a biomechanical point of view, the eyeball possesses very little inertia, thus requiring very small forces to be moved, even at high speed (saccades, the most rapid eye movements, can reach angular speeds of up to 900 °/s). The uniqueness of eye movements is reflected also in the structure of extraocular muscles. Extraocular

muscles have the smallest motor units of all voluntary muscles. Furthermore, they lack the classical muscle stretch receptors that are observed in limb muscles, which deliver movement-related feedback signals to motor circuits.

Eye movements, like any other movements around joints, are affected by the so-called redundancy problem. Eyeballs have 3 degrees of freedom, being able to rotate around 3 perpendicular axes (horizontal, vertical, and torsional axes). For each specific direction of gaze, there are infinite theoretical possible orientations of the eyeball. This constitutes for the nervous system the so-called redundancy problem": any possible combination of eye muscles could be activated to reach the same final direction of gaze. The nervous system solves this redundancy problem by a law first described by F.C. Donders (Donders, 1848). Expressed in modern terms, Donders' law states that, of all possible orientations of the eye for each direction of gaze, the oculomotor system chooses only one specific orientation. Independently from Donders' discovery, the German mathematician J.B. Listing (Listing 1845) proposed that any direction of gaze out of the primary position can be represented by a single rotation of the eyeball around a single axis. Any direction of gaze is defined by a first angle defining the direction of eye movement out of the primary position, and a second angle defining the angle between the rotational axis and the primary eye axis. This is equivalent to stating that all rotational axes representing each gaze direction lie on a flat surface, called "Listing's plane." This hypothesis has been confirmed experimentally many times (e.g., Tweed et al., 1990). The measured thickness of Listing's plane amounts to 1-2 degrees. H. von Helmholtz hypothesized that Listing's law is important because if any given gaze direction could be reached with different eye orientations, the same object would be represented in different ways on the retina for the same direction of gaze, leading to perceptual ambiguities. Starting from this hypothesis, and applying a minimum square error principle to variations in the retinal image of an object as a consequence of eye movements, he was able to generate an equation that is compatible with Listing's law (Helmholtz, 1863). Thus, Donders' and Listing's laws contribute to the perceptual stability of the visual world even in the presence of frequent eye movements. These laws have often been cited as an example of how invariant laws governing the external, physical world, which do not change when the coordinate system is changed, have counterparts in the internal, perceptual world.

From the point of view of their physiological role and neural control, eye movements are usually classified in 5 categories: saccades, smooth pursuit movements, vestibulo-ocular reflexes (VOR), optokinetic reflexes (OKR), and vergence. A detailed review of neural control of eye movements is well beyond the scope of this chapter, but we will address one aspect that is relevant to the present discussion, namely the possibility of a conflict between slow pursuit movements and OKR. Slow pursuit movements serve to stabilize the image of a visually fixated target that is moving over the background. An example would be visually tracking the progress of a football player dribbling the ball while running among other players. During smooth pursuit, the retinal image of the target is sharp, while the background is blurred because it moves over the retina (like a camera picture that looks blurred because the camera was moved during exposure). In general, it is not possible to perform a smooth pursuit eye movement voluntarily in the absence of a target.

(Incidentally, this in sharp contrast to limb movements: we can readily move an arm or leg smoothly in the absence of any target). In contrast, the OKR serves to maintain a stable visual image of a moving scene (e.g., when watching the landscape from the window of a moving train). During smooth pursuit of a target, since the background is moving over the retina, the OKR has to be inhibited, otherwise it would interfere with the pursuit of the target. This situation represents an example of a conflict between a top-down, voluntary control mechanisms (smooth pursuit) and a bottom-up, reflex control mechanism (OKR). Smooth pursuit of an object presupposes allocation of attentional resources toward one of many possible targets (in the example above, the football player dribbling the ball) and requires higher cortical activity. According to modern views, the allocation of visual attention is a function of frontoparietal circuits, in particular the frontal eye field (FEF) and the lateral intraparietal area (LIP), in addition to the superior colliculus (see Bisley, 2011, for a review).

This top-down control also emerges during saccadic eye movements. Saccades are rapid, coordinated eye movements that quickly shift the fovea (the region of the retina of maximal visual acuity, covering less than 1 degree of the visual field) from one point of the visual scene to another, allowing successive high-resolution scanning of different points of the scene ("visual scanning"). There is ongoing discussion as to whether or not saccades are ballistic movements. The term "ballistic" means that saccades are not controlled by a feedback mechanism: whenever a saccade is launched ("point of no return"), if the visual target changes its position, the saccade cannot be corrected, and the fovea lands on the wrong fixation point. Consistent with this view is the abovementioned fact that ocular muscles do not possess muscle spindles, which are considered to be crucial sensors that send feedback movement signals to motor centers, allowing correction of errors during movement execution. During saccades, the eyes can reach very high peak velocities, leading to image blurring on the retina. In fact, vision is actively suppressed (by a top-down mechanism) during saccades to prevent retinal blur, implying that we are functionally almost blind during these movements.

It is noteworthy that saccades often anticipate movement of other body joints: if someone turns while walking, gaze jumps forward in saccades along the trajectory, anticipating trajectory by angles that become larger as the angular velocity of turning increases (Imai et al., 2011). There is also evidence that saccade efficiency influences kinematic performance, as shown by correlations between velocity/timing of the saccades and turning performance in Parkinsonian patients (Lohnes & Earhart, 2011).

8.3 Ocular saccades are guided by top-down as well as bottom-up mechanisms

During viewing of a visual scene, what kind of information guides ocular saccades? Are they bottom-up cues, such as the specific saliencies of individual details of a scene? Or are there top-down mechanisms that instruct the eyes where to move?

It is undeniable that, under certain conditions, bottom-up cues are most important in capturing visual attention in an automatic way, through preattentive mechanisms involving parallel processing of several elements at once. This capacity is important in daily life situations, for example when trying to identify a behaviorally relevant element among many other irrelevant objects (distractors). Thus, an element that differs from distractors along only one dimension (e.g., a red "T" embedded in several blue "Ts") jumps to our attention immediately. Experimentally, it can be shown that the time we need to find the anomalous element is fairly independent from the number of distractors. Conversely, when the anomalous element differs along two or more dimensions (e.g., a red "T" embedded amidst several blue "Ts" and red "Ls") search time increases almost linearly with the number of distractors, which is consistent with a sequential search (see e.g., Julesz and Bergen, 1983). In this latter case, it is difficult to explain the result by a bottom-up, saliency-driven mechanism. Top-down search strategies are most likely at work.

In the 1960s, the Russian physiologist A. Yarbus (1961) performed experiments that have dramatically changed our view of top-down mechanisms controlling gaze. Using rudimentary gaze tracking technology available at that time, Yarbus was able to show that task requirements deeply influence ocular saccades: he demonstrated that, when viewing the same complex picture (e.g., a painting by the Russian painter I. Repin called "They did not expect him"), the pattern of saccades changes dramatically depending on the instructions given by the experimenter to the viewer (e.g., "remember the clothes worn by the people in the picture" versus "estimate how long the unexpected visitor has been away"). Thus, eye movements are also influenced by top-down mechanisms or the search strategy, in addition to bottom-up mechanisms. In fact, people can be made "blind" to obvious elements of a visual scene that would be immediately spotted by a naïve observer by leading them to focus on irrelevant details (the phenomenon of inattentional blindness; see below). With regard to neural mechanisms guiding visual scanning of a scene, it is currently assumed that the path of saccades is determined by maps of attentional priority, which are constructed using a combination of bottom-up and top-down cues. According to this modern view, the frontoparietal network that subserves voluntary eye control plays a crucial a role for the setup of attentional maps (Bisley, 2011).

8.4 Voluntary control of eye movements and shifts of attention

The above brief discussion of the roles of bottom-up and top-down processes in the control of eye movements indicates that they are strongly influenced by the viewer's goals and knowledge states. This link between gaze and underlying mental processes may be one reason why adults rely heavily on direction of gaze and fixational eye movements when trying to make sense of the behavior of others. Thus, in the abovementioned example of eye turning anticipating body turning, one

might guess in advance whether the individual is going to turn left or right by monitoring his gaze.

However, gaze is a very potent cue of intentions, but it is not always a valid cue, as illustrated by two examples to which we now turn. Consider first the common situation in which we find ourselves looking at an object, but with our attention actually directed to some other real or imaginary object. In this case, the visually attended object appears to serve as an "anchor" of the "physical eye," while the "internal eye" is actually directed toward something else. This example illustrates the fact that, although shifts of attention and shifts of visual fixation rely on common mechanisms, they are in fact separable. It is also an illustration of the phenomenon of covert attention, addressed empirically for the first time by Posner (1980). In covert attention tasks, participants direct attention to a position in the space other than a fixation point. By evaluating whether participants notice stimuli that appear at random positions in the visual space, it is possible to map the path taken by covert visual attention (the path of the "internal eye"). Consistent with Posner's results, Richards et al. (2012) used eye tracking technology combined with inattentional blindness tasks to demonstrate that the ability to spot an unexpected stimulus was not contingent upon fixating it, suggesting that some individuals located the stimulus via covert attention mechanisms.

Consider next the observation that direction of gaze can be exploited to conceal the real target of attention. Perhaps one of the best illustrations of this comes from the world of professional football. When shooting penalty kicks, some players tend to look at the goalkeeper while covertly attending to the location where they plan to kick the ball. In this case, the player takes advantage of our tendency to rely on direction of gaze as an indicator of future action to fool the goalkeeper into moving to a particular location in the goal box, all the while simultaneously attending to and planning the shot toward a different location. However, there are limits to this voluntary uncoupling of gaze and attention. For example, Wilson et al. (2009) have shown that even in experienced footballers, anxiety impairs the ability to uncouple gaze from attention while kicking a penalty. In the presence of a "threatening" goalkeeper, footballers made longer fixations on the goalkeeper and their shots became more centralized than they were in the presence of a neutral goalkeeper. Wilson et al. interpret their results as a consequence of an increased influence of the stimulus-driven attentional control system overcoming voluntary, top-down attentional control.

The above discussion regarding the link between fixational eye movements and visual attention leads to the question of whether there are common neural networks for top-down control both of eye movements and attention. This question has only recently been tackled. Recent experiments suggest that the FEF, the oculomotor area of the frontal lobe that is crucial for controlling voluntary saccades, is also an area dedicated to top-down control of visual attention. In fact, experiments in which monkeys were operantly trained through neurofeedback to voluntarily control the activity of neurons within the FEF have shown that operantly driven FEF activity was primarily associated with selective visual attention, and not oculomotor preparation. In this experimental protocol, selective attention correlated with voluntary,

but not spontaneous, fluctuations in FEF activity (Schafer and Moore 2011). Besides pointing to a possible mechanism for uncoupling of fixational eye movements and selective attention, these experimental observations suggest the possibility of using neurofeedback to learn to voluntarily control attention, opening up a potential therapeutic strategy for attentional disorders.

8.5 The central role of inhibitory mechanisms for voluntary control of behavior

It is well known that we can avert gaze if we do not want to look at something. The role of top-down mechanisms in controlling eye movements is particularly evident when one considers the so-called antisaccades, which are ocular saccades away from a visual stimulus that is powerful enough to stimulate a reflex saccade (Everling and Fischer, 1998; Hutton and Ettinger, 2006). Inhibition of reflex saccades is also crucial for maintaining attentional focus on task-relevant elements in the presence of distractors. The antisaccade task has become a popular task in eye movement research: participants are instructed that, after presentation of a peripheral target, say in the right half of the screen, they should look away at the mirror position in the left half of the screen. They must therefore suppress the reflexive urge to look at a visual target that appears suddenly in the peripheral visual field and must instead look in the opposite direction. Performing this task presupposes two different processes: a) the automatic, visual grasp reflex must be suppressed; b) the stimulus vector must be inverted into the saccade vector. The antisaccade task has become an important test for deficits of the prefrontal cortex-basal ganglia loop and the associated deficits of planning for future behavior based on a set of rules. Recent research shows that patients affected by different neurological and psychiatric conditions affecting the frontal lobes or the basal ganglia demonstrate reduced ability to perform the antisaccade task, indicating a deficit in top-down inhibitory mechanisms. In particular, patients with Parkinson's disease have difficulty in performing the antisaccade task. Recent neuroimaging work suggests that the difficulty in performing the antisaccade task is not related to a dysfunction in movement execution, but that there is an impairment of the processes that lead to a preparatory readiness to perform an action (Cameron et al., 2012).

In the complex situations of daily life, inhibitory mechanisms allow us to override an automatic response with an alternative, voluntary response that is more difficult to perform or appears less "intuitive" (Cameron et al., 2012). The alternative voluntary response requires a corresponding "action plan" that manifests itself in the increased activation of areas that are related to voluntary attention and executive function, in particular the prefrontal cortex. This is related to the well-known phenomenon of reality bias, observed in complex situations of daily life, when we tend to act only on the basis of information that is immediately available, rather than on larger scale information. For example, while driving on the

highway, we have the tendency to change lanes as soon as we see that traffic in the other lane is travelling faster, instead of considering the possibility that this effect may be only a local phenomenon. However, switching lanes at a particular point of the highway may reflect an informed strategy, based on our experience that the lane becomes slower at this point on the highway (e.g., because traffic backs up in the lane because of an exit a few kilometers ahead).

The reality bias is also evident in the behavior of young children. Thus, for example, one well-documented phenomenon of the preschool period is that while 3-year-old children consistently fail the standard false belief task used to assess theory of mind, 4-year-olds are successful. In one version of the false belief task, a doll plays with her favorite toy and then places it in a safe location before going out to play. After the doll leaves, the experimenter and the child sneakily move the toy to a new location. The child is then asked where the doll will look for the toy upon her return. While 4-year-olds correctly state that the doll will look in the original location, 3-year-olds consistently say that she will look in the new location (i.e., where the toy is currently). One explanation that has been proposed for this developmental shift is that 3-year-olds are less able to resist the influence of the immediate environment, and that what changes between 3 and 4 is the ability to self-regulate, or step back from the immediate context in order to consider other available information (e.g., Mitchell, 1994). Thus, because the reality of the situation is more salient to 3-year-olds than the beliefs of the doll, it exerts a stronger influence on their response to the question, with the result that their answer aligns with the current reality (i.e., that the doll will look where the toy is now).

8.6 Concluding remarks

In this very brief overview of the basic principles of voluntary control of eye movement, a central theme has been the role of inhibitory mechanisms. Inhibitory mechanisms serve to suppress reflex responses when they are not appropriate and allow selection from among many different possible action plans. Such reflex responses can be elicited by salient external stimuli, such as a light point appearing in the visual field, or by internal stimuli, e.g., when we visually search for something for which we do not exactly remember the position. Loss of inhibitory mechanisms results in the inability to suppress reflexive responses to external or internal stimuli. People with lesions of the prefrontal cortex have difficulty in inhibiting reflex saccades and are also impaired in following an action plan. They tend to alter their plans quickly based on rapid, superficial appraisals. Loss of inhibition also results in involuntary movements that are sometimes extremely unpleasant for the patient, as demonstrated by neurological diseases such as Parkinson disease, Huntington disease, Tourette syndrome, or the alien hand syndrome.

Such inhibitory mechanisms are, in our opinion, a prerequisite of freedom, because they allow us to be in command of our actions instead of being influenced by the present reality that commands our attention. In the classical theory of virtues,

prudence is the virtue that allows one to govern oneself through reason. The word *prudence* stems from the latin word *prudentia*, which is related to *providentia* = foresight. Prudent behavior is therefore behavior that takes in account information that is not available in the here and now ("Wait! Think!").

Inhibitory mechanisms may constitute a neurobiological counterpart of the theory of "virtues," and their negative counterpart, vices. The theory of the so-called cardinal virtues (prudence, temperance, fortitude, justice) has been a centerpiece of Greek moral philosophy and has been further developed in Christian philosophy and moral theology. Virtues can become like a "second nature" of a person, so that doing good and avoiding evil become more and more effortless, but it never becomes automatic. Inhibitory mechanisms are still at work. Even in biographies about Saints who have become famous for their patience and meekness, we read that under certain circumstances they had to suppress their bad temper. For example, Saint Jean-Baptiste-Marie Vianney, the Cure of Ars, was sometimes seen to carry a handkerchief in his hand and to squeeze it strongly when speaking with people that would irritate him.

In everyday life, we usually say "no" many more times than we say "yes," often without being conscious of a moral dilemma. This ability to inhibit behavioral patterns that are inconsistent with our long-term goals gives us the ability to sustain a specific course in life, despite countless stimuli and adverse events that might otherwise let us deviate from our intended course. Such a behavior is typically human and is compatible with the existence of free will.

Acknowledgments Preparation of this chapter was supported by grants from the European Union (231722, IM-CLeVeR) and from the National Institutes of Health (R01 HD054979, R21 HD068584)

References

Bergen, J. R., & Julesz, B. (1983). Parallel versus serial processing in rapid pattern discrimination. *Nature, 303*, 696–698.

Bisley, J. W. (2011). The neural basis of visual attention. *The Journal of Physiology,* 49–57.

Cameron, I. G., et al. (2012). Impaired executive function signals in motor brain regions in Parkinson's disease. *Neuroimage, 60*, 1156–1170.

Donders, F. C. (1848). Beiträge zur Lehre von den Bewegungen des menschlichen Auges. *Holländische Beiträge Anat. Physiol. Wiss I*, 104–145.

Everling, S., & Fischer, B. (1998). The antisaccade: A review of basic research and clinical studies. *Neuropsychologia, 36*, 885–899.

Horowitz, T. S., Wolfe, J. M., Alvarez, G. A., Cohen, M. A., & Kuzmova, Y. I. (2009). The speed of free will. *The Quarterly Journal of Experimental Psychology*, 2262–2288.

Helmholtz, H. V. (1863). Ueber die normalen Bewegungen des menschlichen Auges. *Archiv für Ophthalmologie, 9*(2):153–214.

Hutton, S. B., & Ettinger, U. (2006). The antisaccade task as a research tool in psychopathology: A critical review. *Psychophysiology, 43*, 302–313.

Imai, T., Moore, S. T., Raphan, T., & Cohen, B. (2011). Interaction of the body, head, and eyes during walking and turning. *Experimental Brain Research, 136*, 1–18.

Ingram, J. N., & Wolpert, D. M. (2011). Naturalistic approaches to sensorimotor control. *Progress in Brain Research, 191*, 3–29.

Jonas, H. (2001). The nobility of sight: A study in the phenomenology of the senses. In: *The phenomenon of life. Toward a philosophical biology*. Northwestern University Press.

Mitchell, P. (1994). Realism and early conception of mind. A synthesis of phylogenetic and ontogenetic issues. In C. Lewis & P. Mitchell (Eds.), *Children's early understanding of mind* (pp. 19-46). Hillsdale, NJ: Erlbaum.

Munoz, D. P., & Everling, S. (2004). Look away: The anti-saccade task and the voluntary control of eye movement. *Nature Reviews Neuroscience, 5*, 218–228.

Listing, J. P. (1845). *Beitrag zur Physiologischen Optik*. Goettingen, Vandenhoeck and Ruprecht: Goettinger Studien.

Lohnes, C. A., & Earhart, G. M. (2011). Saccadic eye movements are related to turning performance in Parkinson Disease. *Journal of Parkinson's Disease, 1*, 109–118.

Posner, M. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology, 32*, 3–25.

Richards, A. M., Hannon, E., & Vitkovic, M. (2012). Distracted by distractors: eye movements in a dynamic inattentional blindness task. *Consciousness and Cognition, 21*, 170–176.

Schafer, R. J., & Moore, T. (2011). Selective attention from voluntary control of neurons in prefrontal cortex. *Science, 332*, 1568–1571.

Tweed, D., Cadera, W., & Villis, T. (1990). Computing three-dimensional eye position quaternions and eye velocity from search coil signals. *Vision Research, 30*, 97–110.

Wilson, M. R., Wood, G., & Vine, S. J. (2009). Anxiety, attentional control, and performance impairment in penalty kicks. *Journal of Sport & Exercise Psychology, 31*, 761–775.

Woodward, A. L. (2003). Infants' developing understanding of the link between looker and object. *Developmental Science, 6*(3), 297–311.

Yarbus, A. L. (1961). Eye movements during the examination of complicated objects. *Biofizika, 6*(2), 52–56.

Chapter 9
The Mirror Mechanism as Neurophysiological Basis for Action and Intention Understanding

Leonardo Fogassi and Giacomo Rizzolatti

Abstract Mirror neurons are neurons discovered in the premotor and parietal cortex that become active during observation and execution of motor acts indicating their crucial role in action understanding. There is, however, still controversy about their role in social cognition and its contribution to understanding others' actions and intentions. Recent studies in monkeys and humans have shed light on the properties of the parieto-frontal mirror system and its functional relevance for cognition. We conclude that, although there are several mechanisms through which one can understand other individuals' behavior, the parieto-frontal mirror mechanism is the only one that allows understanding others' actions from the inside and gives the observing individual a first-person person grasp of other individuals' motor goals and intentions.

Keywords Mirror neurons • Goal coding • Parieto-frontal mirror system • First-person knowledge • Autistic patients • Plasticity

L. Fogassi (✉)
Department of Neuroscience, University of Parma, v. Volturno 39, 43100 Parma, Italy

Italian Institute of Technology, Rete Multidisciplinare Tecnologica, University of Parma, B.go Carissimi, 10, 43100 Parma, Italy
e-mail: leonardo.fogassi@unipr.it

G. Rizzolatti
Department of Neuroscience, University of Parma, v. Volturno 39, 43100 Parma, Italy

Italian Institute of Technology, Rete Multidisciplinare Tecnologica, University of Parma, Parma, Italy

Italian Institute of Technology, Brain Center for Social and Motor Cognition, University of Parma, Parma, Italy
e-mail: giacomo.rizzolatti@unipr.it

A. Suarez and P. Adams (eds.), *Is Science Compatible with Free Will?: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience*,
DOI 10.1007/978-1-4614-5212-6_9,

9.1 Introductory Remarks

The mechanism represented by mirror neurons (mirror mechanism) unifies action perception and action execution (Di Pellegrino et al. 1992; Gallese et al. 1996; Rizzolatti et al. 1996a; Gallese et al. 2002; Fogassi et al. 2005; Rozzi et al. 2008). The essence of this mechanism is the following: each time an individual observes another individual performing an action, a set of neurons that encode that action is activated in the observer's own cortical motor system.

In this chapter we will first introduce the basic function of goal coding in the motor system, and describe the properties of the parieto-frontal action observation/action execution (mirror) circuit in monkeys and humans. Then we will show how, based on first-person knowledge, the mirror system encodes the intention of other individuals and how this function can be impaired in autistic patients. Finally we will show some evidence of plasticity within the mirror system.

9.2 Goal Coding in the Monkey Cerebral Cortex

A traditional view on information processing in the cerebral cortex maintained that its posterior (parietal and temporal) sector is devoted to perception (high order elaboration of sensory input), while its anterior (frontal) sector plays a crucial role in movement programming and execution, on the basis of information provided by the "perceptual" part of the cortex. This basically serial view was challenged by the neuroanatomical and neurophysiological data accumulated in the last three decades. Briefly, neuroanatomical data showed that most of the connections between posterior and anterior cortical areas are reciprocal, thus indicating that the flow of information runs in parallel, leading to a strict reciprocal influence between action and perception (Rizzolatti et al. 1998; Rizzolatti and Luppino 2001). Neurophysiological data showed that the motor cortex, far from being a purely executive cortical sector, contains stored representations of the goals of motor acts (Rizzolatti et al. 1988). Through the above-mentioned neuroanatomical connections, the role of these motor signals is that of providing a meaning to the incoming sensory information provided by the posterior cortical areas. For example, when I see an object in the external space, besides visual recognition, its physical properties are immediately transformed in a motor format, that is, in the goal-related motor act appropriate for interacting with that object (see Jeannerod et al. 1995). However, if the context does not allow the execution of this motor act, the activation of the motor system remains in the state of a potential motor act. Thus, our understanding of the external world is, at least partly, based on the automatic activation of the motor system.

Evidence for goal coding in the motor system has been given by single neurons recording experiments carried out on ventral premotor cortex (area F5, see Fig. 9.1) showing that most of its motor neurons discharge during the execution

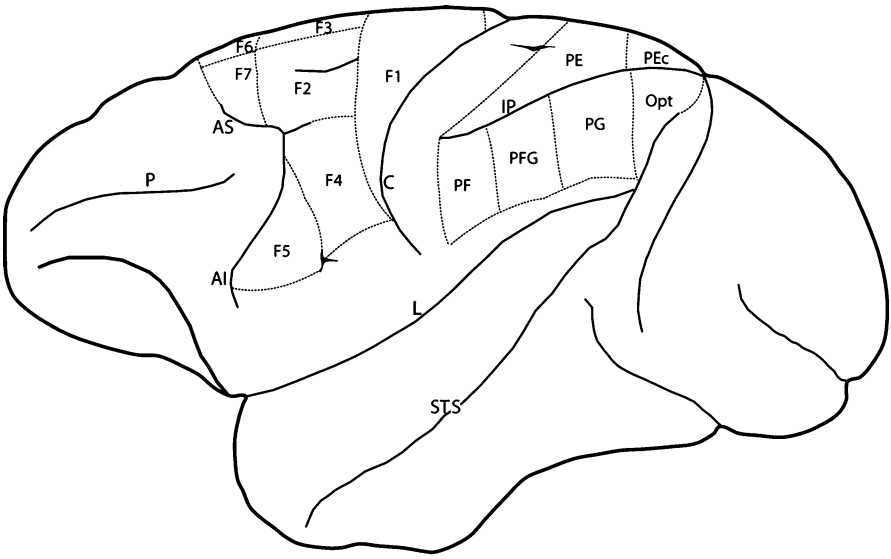


Fig. 9.1 Lateral view of the monkey brain showing the parcellation of the agranular frontal and posterior parietal cortices. Agranular frontal areas are defined according to Matelli et al. (1985, 1991). All posterior parietal areas are defined according to Pandya and Seltzer (1982) and Gregoriou et al. (2006). *AI* inferior arcuate sulcus; *AS* superior arcuate sulcus; *C* central sulcus; *IP* inferior parietal sulcus; *L* lateral fissure; *P* principal sulcus; *STS* superior temporal sulcus

of goal-related motor acts such as grasping, manipulating, breaking, etc., rather than during execution of simple movements, i.e., body-parts displacements without a specific goal (e.g., finger flexion) (Rizzolatti et al. 1988; Kakei et al. 2001). Compelling evidence that this is the case was recently provided by Umiltà et al. (2008). They recorded single neurons in monkeys trained to grasp objects using two different types of pliers: "normal pliers," which require typical grasping movements of the hand, and "reverse pliers," which require hand movements executed in the opposite order (i.e., closing first and then opening the fingers). The results showed that F5 neurons discharged during the same phase of grasping in both conditions, regardless of whether this involved opening or closing of the fingers.

Area F5 belongs to the above-mentioned set of circuits connecting parietal and frontal cortex. Specifically, it is connected with a sector of the inferior parietal lobule (IPL), namely areas PFG (see Fig. 9.1) and AIP (an area buried inside the rostral part of the inferior parietal sulcus). Interestingly, the functional properties of IPL motor neurons seem to be similar to those of F5 neurons, that is, they are active during the execution of goal-directed motor acts rather than the single movements constituting them (Hyvärinen 1982; Sakata et al. 1995; Fogassi et al. 2005; Rozzi et al. 2008).

9.3 The Parieto-Frontal Mirror Circuit

9.3.1 The Monkey Parieto-Frontal Network

The mirror mechanism was originally discovered in the ventral premotor cortex of the macaque monkey (area F5) (Di Pellegrino et al. 1992; Gallese et al. 1996; Rizzolatti et al. 1996a). Single neuron recordings showed that in this area there are neurons that fire both when a monkey *executes* a specific motor act and when it *observes* another individual (either a conspecific or an experimenter) performing the same motor act (mirror neurons, Fig. 9.2). Mirror neurons do not respond to the simple object presentation and do not respond, or respond only weakly, to the observation of the experimenter performing a hand motor act (e.g., grasping) without a target object.

Although the response of most mirror neurons is not influenced by many visual details of the observed act, some of them show specificity for the direction or the space sector in which the act is performed or the hand (left or right) used by the observed agent.

Since their first discovery it has been suggested that mirror neurons have a prominent role in the understanding of the goal of observed motor acts. However, one could

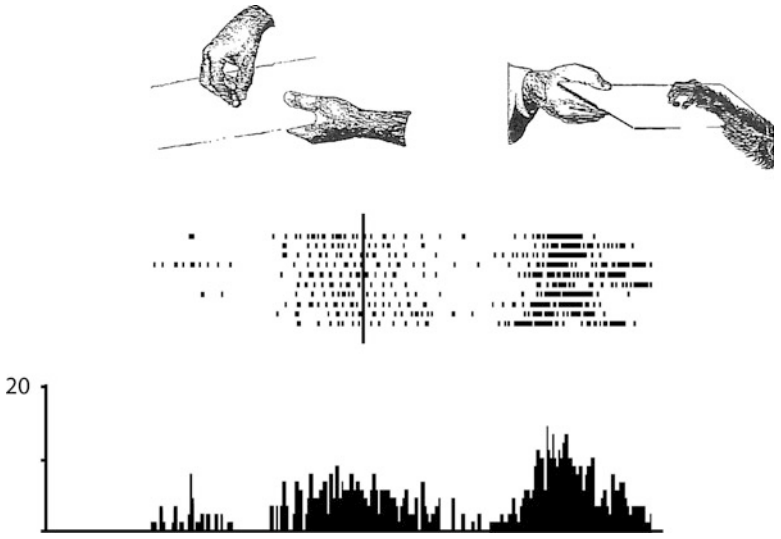


Fig. 9.2 Mirror neuron responding during observation and execution of a hand grasping motor act. The neuron shows a visual response when an experimenter grasps a piece of food in front of the monkey and when the monkey grasps the same piece of food from the experimenter's hand. The silence between the visual and the motor responses corresponds to the time in which the experimenter approaches the plate with food to the monkey, before it grasps it. Rasters and histograms are aligned with the moment in which the experimenter's hand touches the food. Abscissae: time; ordinates: spikes per bin; bin width: 20 ms (modified from Rizzolatti et al. 1996a)

have argued that their response, due to the visual presentation of a motor act, could in principle express a simple visual recognition of a biological movement, without allowing to assign it a motor meaning. This criticism was overcome by two studies in which it was demonstrated that the vision of the motor act is not a necessary requisite for activating mirror neurons. In a first study (Umiltà et al. 2001) it has been demonstrated that mirror neurons responded both when the monkey could fully observe a grasping act and when it could see only part of it because the hand–target interaction was hidden behind a screen. Interestingly, the response was absent when the monkey knew that no object was present behind the screen (mimicked hidden motor act), suggesting that mirror neurons have access to prior contextual information in order to retrieve the motor representation corresponding to the observed motor act, despite the absence of a full visual description of the motor act.

In a second study (Kohler et al. 2002) it has been demonstrated the capacity of mirror neurons to respond to the sound of motor acts. These neurons have been called audio-visual mirror neurons and constitute a subcategory of F5 mirror neurons. They activate when a monkey not only observes, but also hears the sound of a motor act. Their response is specific for the type of motor act seen and heard. For example, these neurons respond to peanuts breaking when the act is only observed, only heard, or both heard and observed, and do not respond to the vision and sound of another act, or to nonspecific sounds. The presence of audio-visual mirror neurons demonstrates that, beyond the visual input, also the acoustic input related to biological actions can have access to the representation of the goal of motor acts.

The most important property of mirror neurons is the congruence they show between the visual and the motor response, that is, the matching between the goal of the observed motor act and that of the executed motor act. This property is crucial, because it enables the observer to understand *what* another individual is doing. In other words, during observation of a motor act, the corresponding motor representation is automatically retrieved in the motor cortex of the observer. Note that, during observation, observers normally do not mimic the observed motor acts. This means that an inhibitory mechanism is at work, so that the "motor resonance" elicited in the observer does not become an overt motor output. Interestingly, very recently Kraskov et al. (2009) demonstrated that half of F5 mirror neurons that activated during grasping execution were inhibited during grasping observation. This inhibition could, at least in part, explain why the observed motor act is not automatically converted in its execution.

Up to now we described, as main function of mirror neurons, that of understanding the goal of motor acts, without entering in the issue of which could be their role in the behavioral reactions consequent to the observation of other individuals' actions. A more recent study (Caggiano et al. 2009) allows to propose some answer in this direction. The main aim of the study was that of assessing whether the discharge of mirror neurons can be modulated by the distance at which the observed act is performed. The same motor act was performed by the experimenter inside the monkey reaching space (peripersonal space) or outside it (extra-personal space). It has been found that 50 % of mirror neurons were differently active in the two conditions. Of them, half discharged stronger when the experimenter grasped a

piece of food within the monkey peripersonal working space, while the other half responded better when the same motor act was performed in the extra-personal space. Interestingly, when the monkey working space was shortened by the presence of a barrier, extra-personal neurons started to discharge strongly also within the peripersonal space, as if it were become far. Taken together, these data suggest that mirror neurons could code other's action within different spaces, and that this property could be related to the possibility to socially interact (cooperate, compete) with others.

Mirror neurons are also present in the rostral part of the IPL, particularly in area PFG (Gallese et al. 2002; Fogassi et al. 2005; Rozzi et al. 2008) and AIP (the anterior intraparietal area) (Rizzolatti et al. 2009). The properties of parietal mirror neurons are quite similar to those of F5. Both areas PFG and AIP are heavily connected with F5 (Borra et al. 2008; Rozzi et al. 2006; Gerbella et al. 2011). These two areas receive higher order visual information from the cortex located inside the superior temporal sulcus (STS) (Rozzi et al. 2006; Borra et al. 2008, see also Fig. 9.1). STS areas encode, as mirror areas, biological actions, but they lack motor properties. AIP receives also connections from the inferior temporal gyrus (Borra et al. 2008). This input could provide the mirror areas with information concerning object identity.

9.3.2 Evidence for New Types of Mirror Neurons

LIP mirror neurons. An interesting function that involves an interaction between two individuals is shared attention. When an individual, for example, is looking in a given direction, an observer located in front of him tends to gaze to the same direction (Gaze following). This behavior can be functional to share the same target at which the first individual's gaze is directed. Neurophysiologically, observation of the eye position of another monkey is known to activate neurons in the STS (Perrett et al. 1992). Only recently, however, it has been demonstrated the presence of mirror neurons for eye movements in the lateral intraparietal area (LIP). This area, located inside the intraparietal sulcus (IPS), is part of a circuit involving the frontal eye field and plays a crucial role in organizing intended eye movements. Most of its neurons discharge when the monkey looks in a specific direction (Barash et al. 1991). Interestingly, a subset of them has been found to discharge also when a monkey observed another monkey looking in the neuron motor preferred direction (Shepherd et al. 2009). This finding suggests that the motor system involved in the control of eye movements towards targets is endowed with a mirror-like mechanism. In sharing attention, the automatic social reaction to another individual's gaze might rely on this mirror-like mechanism.

VIP mirror neurons. Previous studies showed that VIP neurons encode tactile and visual stimuli delivered in the peripersonal space of the monkey (Colby et al. 1993; Duhamel et al. 1998). Ishida et al. (2009) demonstrated that some of these neurons also respond to stimuli presented in the peripersonal space of an individual

located at about one meter far from the monkey and facing it. It is important to keep in mind that area VIP is strictly connected with area F4, which represents peripersonal space and whose neurons discharge during reaching movements. It is plausible, therefore, that neuronal responses that seem to be induced by visual stimuli actually represent potential motor acts directed towards specific body parts (Fogassi et al. 1996). The study on VIP neurons is of great interest because it shows that the mirror mechanism of this area encodes body-directed rather than object-directed motor acts, thus opening fascinating possibilities for individuals to encode the body of others.

Altogether, the described studies on LIP and VIP indicate that the function of mirror neurons is related to the motor properties of the areas in which they are located.

9.4 The Human Parieto-Frontal Mirror System

Sensory, motor, and cognitive functions can be studied in humans by means of electrophysiological (EEG; MEG; TMS) and neuroimaging (PET, fMRI) techniques. These techniques have been successfully employed in the last 15 years to demonstrate that an action observation/action execution mirror circuit also exists in humans.

Brain imaging studies have shown that, as in the monkey, this action observation/action execution mirror circuit is formed by two main regions (Fig. 9.3): (1) the inferior sector of the precentral gyrus plus the posterior part of the inferior frontal gyrus (IFG); (2) the IPL including the cortex located inside the IPS (see Rizzolatti and Craighero 2004; Rizzolatti et al. 2009). Additional cortical areas (such as the dorsal premotor cortex and the superior parietal lobule) have been also occasionally found to be active during action observation. These areas are active when volunteers are asked to observe proximal arm movements directed to a particular location in space (Filimon et al. 2007).

By using single-subject fMRI analyses, evidence has been recently provided that other cortical areas (e.g., SI, SII, middle temporal cortex) become active during action observation and action execution (Gazzola and Keysers 2009). It has been suggested that these activations outside the "classical" mirror areas reflect additional mechanisms (e.g., internal models) that are triggered by the mirror mechanism. These activations would enrich the information about other individuals' actions that the mirror mechanism provides.

In agreement with early findings (Rizzolatti et al. 1996b; Buccino et al. 2001; Decety et al. 2002), a series of new fMRI studies provided strong evidence that the human mirror parieto-frontal circuit encodes the goal of observed motor acts. Gazzola et al. (2007a, b) instructed volunteers to observe video clips in which either a human or a robot arm grasped objects. In spite of differences in shape and kinematics between the human and robot arms, the parieto-frontal mirror circuit was activated in both conditions. These results were extended by Peeters et al. (2009), who investigated the cortical activations in response to the observation of

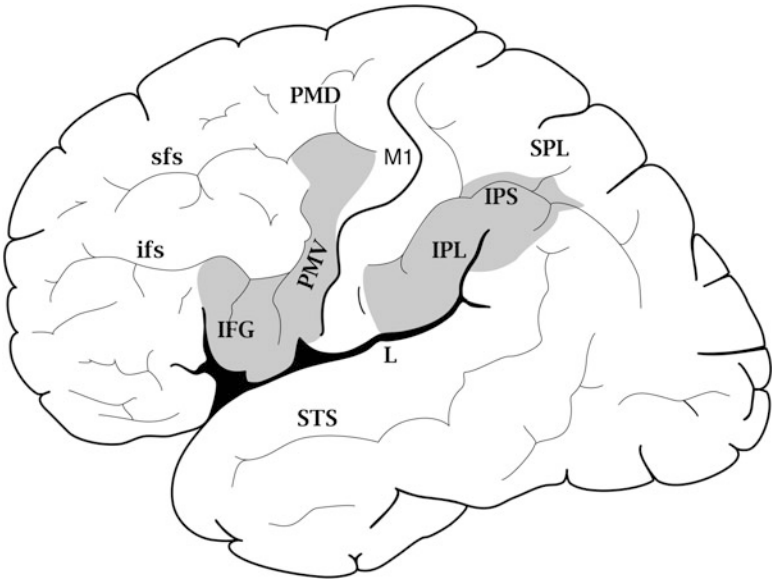


Fig. 9.3 Cortical areas belonging to the parieto-frontal mirror system. Gray-shaded regions indicate cortical sectors activated during action observation, that become also active during execution of the same actions. Note that in some studies additional cortical areas (e.g. dorsal premotor cortex and superior parietal lobule) can activate during observation of reaching or body movements. A rostral sector of the superior temporal sulcus also activate during action observation, but not during action execution. IFG, inferior frontal gyrus; ifs, inferior frontal sulcus; IPL, inferior parietal lobule; IPS, intraparietal sulcus; L, lateral sulcus; M1, primary motor cortex; PMD, dorsal premotor cortex; PMV, ventral premotor cortex; sfs, superior frontal sulcus; SPL, superior parietal lobule; STS, superior temporal sulcus (from Cattaneo and Rizzolatti 2009)

motor acts performed by a human hand, a robot hand, or a tool. They found bilateral activation of a mirror network formed by intraparietal and ventral premotor cortex regardless of the acting effector. In addition, the observation of tool actions produced a specific activation of a rostral sector of the left anterior supramarginal gyrus, suggesting that this sector specifically evolved for tool use.

Unlike monkeys, the parieto-frontal mirror circuit of humans becomes also active during the observation of individual movements (Rizzolatti et al. 1999; Lui et al. 2008). The initial evidence for this mechanism was based on TMS experiments that indicate that the observation of others' movements results in an activation of the muscles involved in the execution of those movements (Fadiga et al. 1995; Strafella and Paus 2000; Gangitano et al. 2001). Additional support comes from EEG and MEG studies showing that the observation of movements without a goal desynchronizes the electroencephalographic rhythms recorded from motor areas (Hari et al. 1998; Cochin et al. 1998; Kilner et al. 2009). These data suggest that in humans both observation of goal-directed actions and of simple movements can activate the motor system. These two types of activation are very likely used for different purposes.

9.5 Understanding Actions Based on First-Person Knowledge

Most of the data reviewed up to now indicate that action understanding is based on a first-person motor knowledge. However, it has been proposed that action understanding could occur by analyzing the different visual elements of the observed actions and applying to them some form of inferential reasoning (see Wood and Hauser 2008). Actually, in some cases, motor behavior might require a mechanism different from mirroring in order to be understood. The capacity of humans to recognize animals' actions that do not belong to the human motor repertoire and cannot be captured by motor generalization is a typical example in this regard. Evidence for the existence of both a mirror and a non-mirror mechanism in non-conspecifics action recognition has been provided by Buccino et al. (2004b). In their fMRI study volunteers were presented with video clips showing motor acts that did or did not belong to the human motor repertoire. The former consisted of ingestive actions performed by a conspecific or by animals (dog and monkey). The latter consisted of communicative gestures (silent speech, dog barking, and monkey lip-smacking). Although all volunteers recognized the observed motor acts regardless of whether or not they belonged to their own motor repertoire, the parieto-frontal mirror system was activated during observation of all ingestive actions and during observation of silent speech. Instead, no activation of parieto-frontal mirror areas was found in the case of those acts that did not belong to it (e.g., dog barking). The areas that became active in the last case were occipital visual and STS areas. These data indicate that the recognition of others' motor behavior can rely on the mere processing of its visual aspects, but it does not provide the observer with information necessary for a real understanding of the message (e.g., the communicative intent of the barking dog). By contrast, when the observed motor act activates the motor system through the mirror mechanism, that action is not only visually recognized but also understood, because there is a sharing of motor goal by the observer and the agent. In other terms, the observed action is understood from the inside in motor terms and not from the outside as a mere visual description.

9.6 Understanding the Motor Intentions of Others

9.6.1 A Matching Mechanism Based on Action Organization

When we perform a complex action we *intend* to achieve a given behavioral goal. Thus, our intention does not correspond to a general preparation to act, but specifies an ultimate goal. In this sense, the term intention is used with a meaning different from that used by other authors in neuroscience. For example, according to some proposals, intention represents a sort of readiness to start a movement, according to others, a preparation of a precise movement or motor act, including programming of

some motor parameters (for example, direction of an impending reaching movement). On the contrary, according to our proposal, the agent's intention includes the selection of a final goal—on the basis of his motivation and of the context—and the organization of the sequence of motor acts necessary to achieve this goal. Interestingly, each motor act belonging to an intentional action has its subgoal, the achievement of which is instrumental for the unfolding of the action sequence, because it prepares the following motor act. The questions are how intentional actions are coded in the parieto-premotor cortical circuits and whether the neurons coding the goal of motor acts are influenced by the ultimate action goal. In order to provide a first answer to this question, grasping neurons were recorded from areas PFG and F5 while the monkey executed a motor task and observed the same task, performed by an experimenter, in which the same motor act (grasping) was embedded into two different actions (grasping to eat and grasping to place) (Fogassi et al. 2005; Bonini et al. 2010). The results showed that a high percentage of parietal and premotor neurons discharged differently when the monkey performed the grasping act, depending on the final goal of the action (either eating or placing) in which the act was embedded. This finding implies that areas F5 and PFG are constituted of chains of neurons in which each neuron encodes a given motor act and is linked to another one selective for the next motor act in the sequence (Fogassi et al. 2005; Rizzolatti et al. 2006). Together they encode a specific action intention (e.g., eating or placing).

Similarly to the motor task, during the visual task it has been found that most mirror neurons discharged differently during observation of grasping, when this act was embedded into different actions. Because in this case the grasping act was performed by the observed agent, it was suggested that the neuronal selectivity for the action goal during grasping observation represents a prediction of the action outcome. Thus, in agreement with the "chain" interpretation of the results of the motor task, the observation of a motor act embedded in an action would activate a chain corresponding to a specific intention. This activation would allow one to understand automatically the motor intentions of others.

These data underline two important concepts. First, the intention to achieve a given motor goal is directly represented in the motor system by a dedicated "chained" neuronal organization. Thus, the motor system does not only encode the goals of motor acts, but also the ultimate action goals. Second, in spite of the mentalistic interpretation of the strategies we use to decode others' intentions, the motor system offers a very simple, automatic mechanism to decode others' intention in most of the contexts of our daily life. Once again, this mechanism provides first-person knowledge of others' behavior.

Evidence that in humans the parieto-frontal mirror circuit is also involved in intention encoding was first provided by an fMRI experiment by Iacoboni et al. (2005). The experiment consisted of three conditions. In the first ("context condition") the volunteers saw video clips showing scenes arranged as to represent an ongoing breakfast or arranged as if the breakfast had just finished ("context" condition); in the second, the volunteers saw video clips showing a hand grasping a mug on an empty background ("action" condition); in the third, they saw videos

showing the same hand motor act within the two contexts ("intention" condition). In this latter condition, the context provided clues for understanding the motor act intention. The results showed that the intention condition induced a stronger activation, relative to the other two conditions, in the caudal IFG of the right hemisphere.

The presence of a chain mechanism underlying intentional actions has been indirectly shown in humans with a behavioral experiment very similar to that used in monkeys and described above. Cattaneo et al. (2007) asked children to grasp a piece of food for eating or for placing it in a container, or to observe an experimenter performing the same actions. During both execution and observation conditions, the EMG activity of the mylohyoid (MH) muscle—a muscle involved in mouth opening—was recorded. Both the execution and the observation of the eating action determined an increase of MH activity during the reaching phase, before object contact, whereas no MH activity was recorded during the execution and the observation of the placing action. This indicates that, as soon as the action starts, the entire motor "chain" involved in action execution is activated. On the observation side, the activation of the same chain would allow the observer to predict what action the agent is going to execute and thus to understand the agent's motor intention.

A mirror mechanism, located in the fronto-mesial areas, can also play a role in representing the motor behavior of others in advance. It has been shown that the "Bereitschaftspotentials," an electrophysiological marker of the readiness to act (Deecke et al. 1969), occurs not only when an individual executes a motor act, but also when the nature and the onset time of an upcoming action performed by another individual is predictable on the basis of a visual cue (Kilner et al. 2004).

9.6.2 Mirroring Intentions and Inferring Reasons

Intention understanding is a multilayered process involving different levels of action representation, from the motor intention that drives a given chain of motor acts to the propositional attitudes (beliefs, desires, etc.) that—at least in humans—can be assumed to explain the observed behavior in terms of its plausible reasons. Thus, while in our daily life we are usually able to understand others' intention through a fast, automatic process, very likely relying on the mirror mechanism, there are cases in which additional inferential processes (Rizzolatti and Sinigaglia 2007; Gallese 2007) are needed. In line with these considerations, recent empirical data showed that, although the parieto-frontal mirror mechanism is active in all conditions in which a motor task has to be directly understood, when volunteers were required to judge the reasons behind the observed actions, there was an activation of a sector of the anterior cingulate cortex and of other areas of the so-called mentalizing network (de Lange et al. 2008). Activation of the same network was also shown in a study (Brass et al. 2007) that investigated unusual actions

performed in implausible vs. plausible contexts, as well as in a study (Liepelt et al. 2008) that studied the neural basis of reason inference in non-stereotypic actions.

The areas belonging to this network have as yet not been demonstrated to have mirror properties. There have been several proposals trying to integrate these two ways of understanding other's intentions (Kilner and Frith 2007; Keysers and Gazzola 2007). However, differently from the mirror system, there are currently no neurophysiological data that can explain how the "mentalizing network" might work.

9.6.3 Intention Understanding in Autistic Patients

Autistic spectrum disorder (ASD) is a syndrome characterized by impairment in social skills, communicative abilities, emotional responses, and motor behavior (see Frith 2003). Although a number of electrophysiological and brain imaging experiments (Oberman et al. 2005; Théoret et al. 2005; Dapretto et al. 2006; Martineau et al. 2008) showed that individuals with ASD have an impairment of the mirror mechanism, some recent behavioral studies have challenged this view (Hamilton et al. 2007; Leighton et al. 2008). Cattaneo et al. (2007) provided an answer to this discrepancy. They asked children with ASD to grasp a piece of food either for eating or for placing it and, in another condition, to observe an experimenter performing these actions. The activity of the mylohyoid (MH) muscle, a muscle involved in mouth opening, was recorded. Unlike typically developing children (see above), in whom MH activation was already present during the "reaching" and "grasping" phases of the grasping-for-eating action, children with ASD showed MH activation only during the "bringing-to-the-mouth" phase. Furthermore, while typically developing children exhibited MH activation when observing a grasping-for-eating action, ASD children did not. These data indicate that children with ASD have a severe impairment in motor organization that includes a deficit in chaining motor acts into intentional actions and, as a consequence, a lack of activation of intentional motor chains during action observation. ASD children, in order to understand others' actions do not use their internal motor knowledge, but another cognitive strategy. This interpretation is supported by a recent study showing that, although ASD children can understand the meaning of a motor act (e.g., grasping) performed by another agent, they are not able to understand the intention of the whole action. In fact, in order to understand intention, they must rely not on the observed motor behavior, but on the semantics of the object that is being manipulated or on the context in which the motor act takes place (Boria et al. 2009).

9.7 Plasticity of the Mirror System

A very important issue strictly linked to the properties of the mirror system is whether mirror neurons activity can be modified by experience and learning. Although in monkeys this issue will be probably best addressed in the future by

chronic recording experiments, some recent monkey data show that mirror neurons can achieve new properties during visuomotor learning. Rochat et al. (2010), in a study in which F5 grasping neurons were recorded in monkeys trained to grasp objects using tools, reported the presence of F5 mirror neurons responding to the observation of grasping motor acts performed by an experimenter with the hand or with a tool, although the response during observation of grasping made with the tool was weaker than that during observation of hand grasping. This study illustrates that when a novel motor act is incorporated in the own motor repertoire, this allows to establish a new motor resonance during the observation of this act, provided that its goal is similar to that achieved with the hand.

Among the studies showing the presence of the mirror system in humans, a couple of them addressed the issue of whether a different motor experience could determine a different activation of this system. In a first fMRI study (Calvo-Merino et al. 2005) participants, who included classical dancers, dancers of Capoeira (a South American dance), and people naïve in professional dance, observed video clips showing steps of classical dance or Capoeira. Although all groups had an activation of the mirror system, the observation of Capoeira with respect to classical dance caused a greater activation in the Capoeira dancers, while the opposite was observed in classical dancers. Naïve subjects did not show differential activation between the two conditions.

In a second study, similar experience-dependent changes in the mirror system have been described in expert (Bangert et al. 2006) and in naïve piano players, but subjected to training (Lahav et al. 2007), that were required to listen to music after motor training.

The plastic change of the mirror system with motor experience was observed in the course of learning by Cross et al. (2006). In their study, expert dancers had to learn and rehearse novel, complex whole-body dance sequences for 5 weeks. Functional MRI was performed every week while the dancers observed and imagined performing movement sequences, half of which were rehearsed and half unpractised. The results showed that the activation of the mirror system was modulated by the dancers' motor experience, with an increase of activity in PMv and IPL during observation of the rehearsed sequences.

The mirror system reveals its plasticity also in situations in which individuals lack an effector or are blind. In a study by Gazzola et al. (2007b) two aplasic individuals, born without arms or hands, participated in an fMRI study in which they had to observe goal-related hand motor acts. Typically developed subjects, observing the same videos, were used as control. In a second part of the study, both aplasic and normal subjects executed mouth and foot motor acts, while only control subjects performed hand motor acts, in order to map the effectors motor representation. This study achieved two important results. First, during observation, aplasic subjects presented a mirror system activation similar to that of controls. Second, during hand motor acts observation, in the frontal cortex they had an activation of the mouth and foot representation. This means that there was a recruitment, from the motor repertoire, of cortical representations involved in the execution of motor acts that achieve similar goals, i.e., taking possession of an object, using different

effectors. Thus, the mirror system is not only modified by motor experience, but also undergoes plastic changes similar to those already demonstrated in sensory systems after deprivation of the afferent input.

In another study, Ricciardi et al. (2009) showed that when congenitally blind patients listened to the sound of actions there is an activation of a fronto-parieto-temporal system, corresponding to the regions activated in the normally sighted controls, during observation of and listening to the same actions. Furthermore, the sound of familiar actions caused a greater activation of the mirror system in both blind and normally sighted subjects. Thus, regions that in normal developing individuals are devoted to visuomotor integration during observation/execution of actions, accomplish the same functions in congenitally blind individuals, by exploiting a different sensory channel.

9.8 Conclusions

The discovery of mirror neurons has opened a wide spectrum of investigations in the motor cognitive domain and beyond, because it constitutes a basic mechanism matching action execution and action observation that allows the understanding of other's actions from inside. Interestingly, this mechanism seems to be a very basic way of understanding, since its presence has been demonstrated not only in humans and monkeys, but also in singing birds, like swamp sparrows (Prather et al. 2008) and zebra finches (Keller and Hahnloser 2009). Furthermore, this mechanism appears to constitute a deep link between individuals that is fundamental for establishing interindividual relationships. The evidence on people with autism suggests a strong role of motor knowledge and of the mirror mechanism, based on this knowledge, in mediating the capacity to understand others' behavior and to entertain interindividual interactions.

Acknowledgement We thank D. Mallamo for his help in preparing illustrations.

References

Bangert, M., Peschel, T., Schlaug, G., Rotte, M., Drescher, D., Hinrichs, H., Heinze, H. J., & Altenmuller, E. (2006). Shared networks for auditory and motor processing in professional pianists: Evidence from fMRI conjunction. *NeuroImage, 30*, 917–926.

Barash, S., Bracewell, R. M., Fogassi, L., Gnadt, J., & Andersen, R. A. (1991). Saccade-related activity in the lateral intraparietal area. II. Spatial properties. *Journal of Neurophysiology, 66*, 1109–1124.

Bonini, L., Rozzi, S., Serventi, F. U., Simone, L., Ferrari, P. F., & Fogassi, L. (2010). Ventral premotor and inferior parietal cortices make distinct contribution to action organization and intention understanding. *Cerebral Cortex, 20*, 1372–1385.

Boria, S., Fabbri-Destro, M., Cattaneo, L., Sparaci, L., Sinigaglia, C., Santelli, E., Cossu, G., & Rizzolatti, G. (2009). Intention understanding in autism. *PloS One, 4*(5), e5596.

Borra, E., Belmalih, A., Calzavara, R., Gerbella, M., Murata, A., Rozzi, S., & Luppino, G. (2008). Cortical connections of the macaque anterior intraparietal (AIP) area. *Cerebral Cortex, 18*, 1094–1111.

Brass, M., Schmitt, R. M., Spengler, S., & Gergely, G. (2007). Investigating action understanding: Inferential processes versus action simulation. *Current Biology, 17*, 2117–2121.

Buccino, G., Binkofski, F., Fink, G. R., Fadiga, L., Fogassi, L., Gallese, V., Seitz, R. J., Zilles, K., Rizzolatti, G., & Freund, H.-J. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: An fMRI study. *The European Journal of Neuroscience, 13*, 400–404.

Buccino, G., Lui, F., Canessa, N., Patteri, I., Lagravinese, G., Benuzzi, F., Porro, C. A., & Rizzolatti, G. (2004). Neural circuits involved in the recognition of actions performed by nonconspecifics: An fMRI study. *Journal of Cognitive Neuroscience, 16*, 114–126.

Caggiano, V., Fogassi, L., Rizzolatti, G., Thier, P., & Casile, A. (2009). Mirror neurons differentially encode the peripersonal and extrapersonal space of monkeys. *Science, 324*, 403–406.

Calvo-Merino, B., Glaser, D. E., Grezes, J., Passingham, R. E., & Haggard, P. (2005). Action observation and acquired motor skills: An FMRI study with expert dancers. *Cerebral Cortex, 15*, 1243–1249.

Cattaneo, L., & Rizzolatti, G. (2009). The mirror neuron system. *Archives of Neurology, 66*, 557–560.

Cattaneo, L., Fabbi-Destro, M., Boria, S., Pieraccini, C., Monti, A., Cossu, G., & Rizzolatti, G. (2007). Impairment of actions chains in autism and its possible role in intention understanding. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 17825–17830.

Cochin, S., Barthelemy, C., Lejeune, B., Roux, S., & Martineau, J. (1998). Perception of motion and qEEG activity in human adults. *Electroencephalography and Clinical Neurophysiology, 107*, 287–295.

Colby, C. L., Duhamel, J.-R., & Goldberg, M. E. (1993). Ventral intraparietal area of the macaque: Anatomic location and visual response properties. *Journal of Neurophysiology, 69*, 902–914.

Cross, E. S., Hamilton, A. F., & Grafton, S. T. (2006). Building a motor simulation de novo: Observation of dance by dancers. *NeuroImage, 31*, 1257–1267.

Dapretto, M., Davies, M. S., Pfeifer, J. H., Scott, A. A., Sigman, M., Bookheimer, S. Y., & Iacoboni, M. (2006). Understanding emotions in others: Mirror neuron dysfunction in children with autism spectrum disorders. *Nature Neuroscience, 9*, 28–30.

Decety, J., Chaminade, T., Grezes, J., & Meltzoff, A. N. (2002). A PET exploration of the neural mechanism involved in reciprocal imitation. *NeuroImage, 15*, 265–272.

Deecke, L., Scheid, P., & Kornhuber, H. H. (1969). Distribution of readiness potential, pre-motion positivity and motor potential of the human cerebral cortex preceding voluntary finger movement. *Experimental Brain Research, 7*, 158–168.

de Lange, F. P., Spronk, M., Willems, R. M., Toni, I., & Bekkering, H. (2008). Complementary systems for understanding action intentions. *Current Biology, 18*, 454–457.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research, 91*, 176–180.

Duhamel, J.-R., Colby, C. L., & Goldberg, M. E. (1998). Ventral intraparietal area of the macaque: Congruent visual and somatic response properties. *Journal of Neurophysiology, 79*, 126–136.

Fadiga, L., Fogassi, L., Pavesi, G., & Rizzolatti, G. (1995). Motor facilitation during action observation: A magnetic stimulation study. *Journal of Neurophysiology, 73*, 2608–2611.

Filimon, F., Nelson, J. D., Hagler, D. J., & Sereno, M. L. (2007). Human cortical representations for reaching: Mirror neurons for execution, observation, and imagery. *NeuroImage, 37*, 1315–1328.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., & Rizzolatti, G. (2005). Parietal lobe: From action organization to intention understanding. *Science, 302*, 662–667.

Fogassi, L., Gallese, V., Fadiga, L., Luppino, G., Matelli, M., & Rizzolatti, G. (1996). Coding of peripersonal space in inferior premotor cortex (area F4). *Journal of Neurophysiology, 76*, 141–157.

Frith, U. (2003). *Autism. Explaining the enigma* (2nd ed.). Oxford: Blackwell Publishing.

Gallese, V. (2007). Before and below Theory of mind: Embodied simulation and the neural correlates of social cognition. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 362*, 659–669.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (2002). Action representation and the inferior parietal lobule. In W. Prinz & B. Hommel (Eds.), *Common mechanisms in perception and action: Attention and performance* (Vol. XIX, pp. 334–355). Oxford: Oxford University Press.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain, 119*, 593–609.

Gangitano, M., Mottaghy, F. M., & Pascual-Leone, A. (2001). Phase-specific modulation of cortical motor output during movement observation. *NeuroReport, 12*, 1489–1492.

Gazzola, V., & Keysers, C. (2009). The observation and execution of actions share motor and somatosensory voxels in all tested subjects: Single subject analyses of unsmoothed fMRI data. *Cerebral Cortex, 19*, 1239–1255.

Gazzola, V., Rizzolatti, G., Wicker, B., & Keysers, C. (2007a). The anthropomorphic brain: The mirror neuron system responds to human and robotic actions. *NeuroImage, 35*, 1674–1684.

Gazzola, V., van der Worp, H., Mulder, T., Wicker, B., Rizzolatti, G., & Keysers, C. (2007b). Aplasics born without hands mirror the goal of hand actions with their feet. *Current Biology, 17*, 1235–1240.

Gerbella, M., Belmalih, A., Borra, E., Rozzi, S., & Luppino, G. (2011). Cortical connections of the anterior (F5a) subdivision of the macaque ventral premotor area F5. *Brain Structure & Function, 216*, 43–65. Epub 2010 Dec 5.

Gregoriou, G. G., Borra, E., Matelli, M., & Luppino, G. (2006). Architectonic organization of the inferior parietal convexity of the macaque monkey. *The Journal of Comparative Neurology, 496*, 422–451.

Hamilton, A. F., Brindley, R. M., & Frith, U. (2007). Imitation and action understanding in autistic spectrum disorders: How valid is the hypothesis of a deficit in the mirror neuron system? *Neuropsychologia, 45*, 1859–1868.

Hari, R., Forss, N., Avikainen, S., Kirveskari, E., Salenius, S., & Rizzolatti, G. (1998). Activation of human primary motor cortex during action observation: A neuromagnetic study. *Proceedings of the National Academy of Sciences of the United States of America, 95*, 15061–15065.

Hyvärinen, J. (1982). Posterior parietal lobe of the primate brain. *Physiological Reviews, 62*, 1060–1129.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the intentions of others with one's owns mirror neuron system. *PLoS Biology, 3*, 529–535.

Ishida, H., Nakajima, K., Inase, M., & Murata, A. (2009). Shared mapping of own and others' bodies in visuotactile bimodal area of monkey parietal cortex. *Journal of Cognitive Neuroscience, 22*, 83–96.

Jeannerod, M., Arbib, M. A., Rizzolatti, G., & Sakata, H. (1995). Grasping objects: The cortical mechanisms of visuomotor transformation. *Trends in Neurosciences, 18*, 314–320.

Kakei, S., Hoffman, D. S., & Strick, P. L. (2001). Direction of action is represented in the ventral premotor cortex. *Nature Neuroscience, 4*, 1020–1025.

Keller, G. B., & Hahnloser, R. H. (2009). Neural processing of auditory feedback during vocal practice in a songbird. *Nature, 457*, 187–190.

Keysers, C., & Gazzola, V. (2007). Integration simulation and theory of mind: From self to social cognition. *Trends in Cognitive Sciences, 11*, 194–196.

Kilner, J. M., & Frith, C. (2007). Action observation: Inferring intentions without mirror neurons. *Current Biology, 18*, R32–R33.

Kilner, J. M., Marchant, J. L., & Frith, C. D. (2009). Relationship between activity in human primary motor cortex during action observation and the mirror neuron system. *PLoS One, 4*, e4925.

Kilner, J. M., Vargas, C., Duval, S., Blakemore, S.-J., & Sirigu, A. (2004). Motor activation prior to observation of a predicted movement. *Nature Neuroscience, 7*, 1299–1301.

Kohler, E., Keysers, C., Umiltà, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science, 297*, 846–848.

Kraskov, A., Dancause, N., Quallo, M. M., Shepherd, S., & Lemon, R. N. (2009). Corticospinal neurons in macaque ventral premotor cortex with mirror properties: A potential mechanism for action suppression? *Neuron, 64*, 922–930.

Lahav, A., Saltzman, E., & Schlaug, G. (2007). Action representation of sound: Audiomotor recognition network while listening to newly acquired actions. *The Journal of Neuroscience, 27*, 308–314.

Leighton, J., Bird, G., Charman, T., & Heyes, C. (2008). Weak imitative performance is not due to a functional 'mirroring' deficit in adults with Autism Spectrum Disorders. *Neuropsychologia, 46*, 1041–1049.

Liepelt, R., Von Cramon, D. Y., & Brass, M. (2008). How do we infer other's goals from non stereotypic actions? The outcome of context-sensitive inferential processing in right inferior parietal and posterior temporal cortex. *NeuroImage, 43*, 784–792.

Lui, F., Buccino, G., Duzzi, D., Benuzzi, F., Crisi, G., Baraldi, P., et al. (2008). Neural substrates for observing and imagining non-object-directed actions. In C. Keysers & L. Fadiga (Eds.), *The mirror neuron system (special issue of Social Neurosci.)* (pp. 261–275). New York: Psychol. Press.

Martineau, J., Cochin, S., Magne, R., & Barthelemy, C. (2008). Impaired cortical activation in autistic children: Is the mirror neuron system involved? *International Journal of Psychophysiology, 68*, 35–40.

Matelli, M., Luppino, G., & Rizzolatti, G. (1985). Patterns of cytochrome oxidase activity in the frontal agranular cortex of the macaque monkey. *Behavioural Brain Research, 18*, 125–136.

Matelli, M., Luppino, G., & Rizzolatti, G. (1991). Architecture of superior and mesial area 6 and the adjacent cingulate cortex in the macaque monkey. *The Journal of Comparative Neurology, 311*, 445–462.

Oberman, L. M., Hubbard, E. M., McCleery, J. P., Altschuler, E. L., Ramachandran, V. S., & Pineda, J. A. (2005). EEG evidence for mirror neuron dysfunction in autism spectrum disorders. *Brain Research. Cognitive Brain Research, 24*, 190–198.

Pandya, D. N., & Seltzer, B. (1982). Intrinsic connections and architectonics of posterior parietal cortex in the rhesus monkey. *The Journal of Comparative Neurology, 204*, 204–210.

Peeters, R., Simone, L., Nelissen, K., Fabbri-Destro, M., Vanduffel, W., Rizzolatti, G., & Orban, G. A. (2009). The representation of tool use in humans and monkeys: Common and unique human features. *The Journal of Neuroscience, 29*, 11523–11539.

Perrett, D. I., Hietanen, J. K., Oram, M. W., & Benson, P. J. (1992). Organization and functions of cells responsive to faces in the temporal cortex. *Philosophical Transactions of the Royal Society of London, 335*, 23–30.

Prather, J. F., Peters, S., Nowicki, S., & Mooney, R. (2008). Precise auditory-vocal mirroring in neurons for learned vocal communication. *Nature, 451*, 249–250.

Ricciardi, E., Bonino, D., Sani, L., Vecchi, T., Guazzelli, M., Haxby, J. V., Fadiga, L., & Pietrini, P. (2009). Do we really need vision? How blind people "see" the actions of others. *The Journal of Neuroscience, 29*, 9719–9724.

Rizzolatti, G., & Craighero, L. (2004). The mirror neuron system. *Annual Review of Neuroscience, 27*, 169–192.

Rizzolatti, G., & Luppino, G. (2001). The cortical motor system. *Neuron, 31*, 889–901.

Rizzolatti, G., & Sinigaglia, C. (2007). Mirror neurons and motor intentionality. *Functional Neurology, 22*, 205–210.

Rizzolatti, G., Fadiga, L., Fogassi, L., & Gallese, V. (1999). Resonance behaviors and mirror neurons. *Archives Italiennes de Biologie, 137*, 85–100.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2009). The mirror neuron system: A motor-based mechanism for action and intention understanding. In M. Gazzaniga (Ed.), *The cognitive neuroscience* (pp. 625–640). Cambridge: MIT Press.

Rizzolatti, G., Luppino, G., & Matelli, M. (1998). The organization of the cortical motor system: New concepts. *Electroencephalography and Clinical Neurophysiology, 106*, 283–296.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996a). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3*, 131–141.

Rizzolatti, G., Ferrari, P. F., Rozzi, S., & Fogassi, L. (2006). The inferior parietal lobule: Where action becomes perception. *Novartis Foundation Symposium, 270*, 129–140.

Rizzolatti, G., Camarda, R., Fogassi, M., Gentilucci, M., Luppino, G., & Matelli, M. (1988). Functional organization of inferior area 6 in the macaque monkey: II. Area F5 and the control of distal movements. *Experimental Brain Research, 71*, 491–507.

Rizzolatti, G., Fadiga, L., Matelli, M., Bettinardi, V., Paulesu, E., Perani, D., & Fazio, F. (1996b). Localization of grasp representation in humans by PET: Observation versus execution. *Experimental Brain Research, 11*, 246–252.

Rochat, M. J., Caruana, F., Jezzini, A., Escola, L., Intskirveli, I., Grammont, F., Gallese, V., Rizzolatti, G., & Umiltà, M. A. (2010). Responses of mirror neurons in area F5 to hand and tool grasping observation. *Experimental Brain Research, 204*, 605–616.

Rozzi, S., Calzavara, R., Belmalih, A., Borra, E., Gregoriou, G. G., Matelli, M., & Luppino, G. (2006). Cortical connections of the inferior parietal cortical convexity of the macaque monkey. *Cerebral Cortex, 16*, 1389–1417.

Rozzi, S., Ferrari, P. F., Bonini, L., Rizzolatti, G., & Fogassi, L. (2008). Functional organization of inferior parietal lobule convexity in the macaque monkey: Electrophysiological characterization of motor, sensory and mirror responses and their correlation with cytoarchitectonic areas. *The European Journal of Neuroscience, 28*, 1569–1588.

Sakata, H., Taira, M., Murata, A., & Mine, S. (1995). Neural mechanisms of visual guidance of hand action in the parietal cortex of the monkey. *Cerebral Cortex, 5*, 429–438.

Shepherd, S. V., Klein, J. T., Deaner, R. O., & Platt, M. L. (2009). Mirroring of attention by neurons in macaque parietal cortex. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 9489–9494.

Strafella, A. P., & Paus, T. (2000). Modulation of cortical excitability during action observation: A transcranial magnetic stimulation study. *NeuroReport, 11*, 2289–2292.

Théoret, H., Halligan, E., Kobayashi, M., Fregni, F., Tager-Flusberg, H., & Pascual-Leone, A. (2005). Impaired motor facilitation during action observation in individuals with autism spectrum disorder. *Current Biology, 15*, R84–R85.

Umiltà, M. A., Escola, L., Intskirveli, I., Grammont, F., Rochat, M., Caruana, F., Jezzini, A., Gallese, V., & Rizzolatti, G. (2008). How pliers become fingers in the monkey motor system. *Proceedings of the National Academy of Sciences of the United States of America, 105*, 2209–2213.

Umiltà, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., & Rizzolatti, G. (2001). "I know what you are doing": A neurophysiological study. *Neuron, 32*, 91–101.

Wood, J. N., & Hauser, M. D. (2008). Action comprehension in non-human primates: Motor simulation or inferential reasoning*? Trends in Cognitive Sciences, 12*, 461–465.

Chapter 10
On the Quest for Consciousness in Vegetative State Patients Through Electrical Neuroimaging

S.L. Gonzalez, S. Perrig, and R. Grave de Peralta

Abstract Consciousness remains an ill-defined concept. This is reflected in clinical practice as there is no objective way to determine that an unresponsive patient is aware of himself and his/her surroundings. However, from the correct answer to this question depends the diagnosis and eventually the continuation of life sustaining aid. Here we discuss how to build on top of recent progress in the field of reverse neural engineering to implement a Test able to detect objective markers of consciousness for completely unresponsive patients. By focusing on the so-called "soft problem of consciousness"—the correlation between the brain and mental functions—we briefly sketch how we plan to provide partial answers the following questions: (1) What are the necessary conditions to confirm that a conscious mind is enclosed in a completely paralyzed body?, (2) How can we extract these responses from neural activity alone?, (3) How could these signals be exploited to establish a minimal dialogue between the patient and a physician using a system that interprets the neural responses?, (4) Is awareness localized to certain neural structures or, instead, is it a global process that depends on the activation of a critical mass of neurons?

S.L. Gonzalez
Department of Clinical Neuroscience, Electrical Neuroimaging Group,
Geneva Neuroscience Center, University of Geneva. NEUCLI, Rue Gabrielle-Perret-Gentil, 4, 1211, Geneva 14, Switzerland
e-mail: Sara.GonzalezAndino@electrical-neuroimaging.ch

S. Perrig
Sleep Research Lab. and Geneva Neuroscience Center, Service of Neurospychiatry, Geneva University Hospital, Ch. du Petit Bel-Air, 2, 1255 Chene Bourg, Geneva, Switzerland
e-mail: Stephen.Perrig@hcuge.ch

R. Grave de Peralta (✉)
Department of Clinical Neuroscience, Electrical Neuroimaging Group, Sleep Research Lab. and Geneva Neuroscience Center, Geneva University Hospital, NEUCLI,
Rue Gabrielle-Perret-Gentil, 4, 1211, Geneva 14, Switzerland
e-mail: Rolando.Grave@electrical-neuroimaging.ch

A. Suarez and P. Adams (eds.), *Is Science Compatible with Free Will?: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience*,
DOI 10.1007/978-1-4614-5212-6_10,

Keywords Consciousness • Vegetative states • Electrical neuroimaging • Awareness • EEG

10.1 Introduction: Consciousness Seen from the Clinical Perspective

While lacking a single definition, consciousness is often associated with awareness. As expressed by Schneider and Velmans, (Schneider, Schneider and Velmans 2007), "Anything that we are aware of at a given moment forms part of our consciousness, making conscious experience at once the most familiar and most mysterious aspect of our lives." A common aspect, shared by all definitions, is that consciousness is an internal attribute that is independent of our will or capacity to produce overt responses. However, insofar as clinical medicine is concerned, "measuring" consciousness hitherto requires the patient's will and capacity to produce overt responses. Then, the principle put forward by Descartes', "Cogito ergo sum," does not seems sufficient, as in medical and legal practice letting others know that we are conscious is of uppermost importance.

Consciousness in medicine (Overgaard 2009; Owen et al. 2009) is assessed by observing a patient's alertness and responsiveness, and can be seen as a continuum of states ranging from alert, oriented to time and place, and communicative, through disorientation, then delirium, then loss of any meaningful communication, and ending with loss of movement in response to painful stimulation (Laureys et al. 2004). After severe brain injury, patients are classified as (1) brain dead, (2) coma, (3) vegetative state, (4) minimally conscious states or (5) locked-in, according to some tests that always involve the existence of purposeful behavior. In this sense, the existent clinical tests for detection of consciousness agree more with the postulates of some philosophers such as F. Brentano (http://en.wikipedia.org/wiki/Franz_Brentano) who suggest that intentionality or aboutness (consciousness is about something) should appear in the definition. While within the philosophy of mind there is no consensus on whether intentionality is a requirement for consciousness, it is clear that motor actions are not a requirement. It is therefore important to develop objectives measures of consciousness that are independent of the subject's possibilities to perform specific actions.

Patients in persistent vegetative states (VS) represent one of the biggest ethical dilemmas in current medical practice. It is nowadays nearly impossible to insure that a patient who has lost all expressive capabilities is no longer conscious unless we measure consciousness differently. It is indeed possible that a patient is consciously aware of self and surroundings, but unable to communicate it. This is apparently not rare given the relatively large proportion of misdiagnosis reported in the literature (Andrews et al. 1996; Childs et al. 1993; Laureys et al. 2004; Owen et al. 2009). However, misdiagnosing a conscious patient as in the persistent vegetative state might have severe implications as in the best cases no rehabilitation

will be attempted but in the worst cases ethical decisions such as terminating life-sustaining treatment might be evoked and undertaken.

We here discuss our approach to try to determine if a non-responsive patient is conscious or not. This problem is indeed similar to the Turing Test that has to determine if the interlocutor is a machine or a human from the responses it gives to the examiner. In the application of a Turing Test to the problem of detecting consciousness in patients the examiner has at his/her disposal just the neural signals that are coming from the patient brain rather than overt verbal responses. In this test, it is assumed that consciousness emerges from the brain—we will deal with the soft problem of consciousness—and as such neural activity is a carrier of information. We expect that neural signals reflect attempted responses from the patient even though overt responses are impossible. For our purposes, some of questions related to the hard problem of understanding the emergence of consciousness are secondary but could be potentially clarified by studying and establishing dialogues with patients where consciousness is altered. We therefore think that by measuring consciousness more objectively we might eventually help to tighten its definition.

In what follows we focus on aspects that are relevant to the implementation of a Patient Machine Interface implementing the Turing Test (PMI-TT) by reverse neural engineering, namely: (1) which neural signals are more appropriate and sensitive to detect consciousness, (2) what is the right battery of questions to be asked by the examiner during the Turing test, and finally (3) How reverse neural engineering extracts information from neural signals alone and take decisions.

10.2 Brain Electrical Activity as an Information Rich Signal

The idea of our PMI-TT to detect consciousness is schematically depicted in Fig. 10.1. In the classical Turing Test—developed to detect the "intelligence" of machines—Player C, the examiner, is tasked with trying to determine which player—A or B—is a computer and which is a human, based on the responses that A and B provide to his questions. In the PMI-TT described here, we want to determine if the patient is in the state A (conscious but unresponsive) or B (unconscious) from the analysis of the neural signals that are recorded in response to the prompts of the examiner C. In effect, the PMI-TT is composed of two loops. In the first loop the examiner (e.g., the physician) uses a Brain Computer Interface (BCI) that presents stimuli, collects neural signals and analyzes them to learn patterns that help to "decode" the responses (e.g., YES or NO) of the patient. Ultimately, deciding if the patient is in state A or B is a binary classification problem repeatedly addressed by reverse neural engineering in the field of BCIs. The second loop evaluates the reliability (R) of the dialogue, i.e., computes the probability of having these answers by chance based on the prompts whose answers are a priori known (e.g., Is your name XXX?). The probability obtained from the second loop is used as a scale to measure residual consciousness. Unreliable dialogues are associated with more altered states of consciousness and this feedback is provided to the examiner.

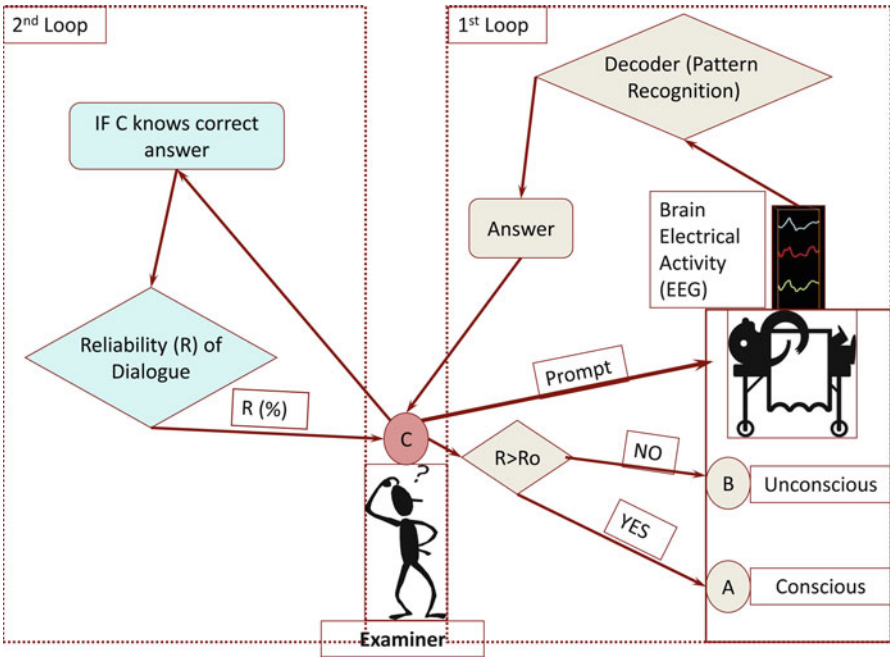


Fig. 10.1 A Patient Machine Interface implementing the Turing Test to assess consciousness: The examiner C (physician) wants to determine if the patient is on the state A (conscious but unresponsive) or B (unconscious) from the analysis of the neural signals that are recorded in response to his prompts. The test is composed of two loops. In the first loop the examiner uses a Brain Computer Interface (BCI) that present stimuli, collect neural signals and analyze them to learn patterns that help to "decode" the responses (e.g. YES or NO) of the patient. The second loop evaluates the reliability of the dialogue which is used as a scale to measure residual consciousness

In short, we want to engage in a Turing-like test procedure with unresponsive (e.g., brain damaged) patients using only their brain signals as carriers of information. Thus, a condition to create an effective PMI-TT is that signals are contingent on external prompts, and react "quickly" to these prompts. This condition poses some constraints on the type of neuroimaging techniques that can be used for the Turing test. Indeed, neuroimaging modalities with high temporal resolution (e.g., electroencephalography ,EEG or magnetoencephalography, MEG) should be preferred to modalities providing coarse temporal resolution (fMRI or PET). Temporal resolution is essential to detect the contingency between the external prompts given by the examiner and the neural responses to it. We also need to obtain responses to each prompt if we want later on to engage in a dialogue with the patient. This implies that neural signals are necessarily very noisy as they come from single trials and therefore special signal processing techniques are required. Some of these responses are assessed (second loop) to get an accumulated measure that can be interpreted as a quantitative measure of consciousness of the subject.

By changing the questions and properly assessing the responses, the system could provide not only a quantitative evaluation of the conscious level but also of the cognitive state.

10.3 Asking the Right Questions During the Turing Test: Awareness of the Self and the Self-Environment

Severe brain damage not only leads to disorders in consciousness but it might affect the capacity of the patients to perceive sensory information. In practical terms there is no sense in engaging in a Turing Test with a patient using questions presented in the screen if there is damage to the visual pathway and vision is impaired. Therefore, it is very important to create a battery of tests that probe the different sensory modalities in patients before attempting to go beyond into the Turing Test.

EEG signals might also be helpful in this direction. The averages over several repetitions of similar stimuli time-locked to stimulus onset are referred in the EEG literature as Event Related Potentials (ERPs) (Regan 1989). ERPs are typically composed of peaks and valleys and can be subdivided according to the latency (time elapsed from stimulus onset to the onset of the peak or valley) into early and late components. It is well established that the early, short-latency components (several milliseconds to several tens of milliseconds after stimulus), reflect propagation of sensory signals from receptors via ascending pathways to the cortex. They have been widely used in clinical neurophysiology to assess if sensory pathways are intact (Regan 1989).

The late ERPs components that appear between 100 and 1,000 ms after the stimulus are of cortical origin and reflect the steps in the cognitive processing of the stimulus. There is a whole body of literature showing how to use ERPs to study different cognitive variables ranging from the processing of physical stimulus features to the processing of semantic stimulus features (and therefore language comprehension).

To test the integrity of the sensory pathways in VS patients we will rely—during the first loop of the PMI-TT—on ERPs. We will assess the visual, auditory and somato-sensory modalities as some other pathways such as olfactory, gustatory, propioceptive, or thermo-algesic are more difficult to test quantitatively at the bed side and are of little interest for establishing a dialogue.

Visual pathway: this can be tested by applying goggles equipped with diodes. While the VS patient is unlikely to be able to fix a target when prompted, if some circuits of visual system are not severely damaged, we should be able to observe visual ERP responses (VEP) at the occipital cortex in response to luminous flashes. In particular, VEP peaks around 100 ms should be recorded in the back of the head, even if fluctuations in latency, amplitude and precise localization are possible.

Auditory pathway: this pathway is easier to test. We use first simple auditory clicks. The evoked potentials can be divided into short latency called Brainstem

Auditory Evoked Potentials (BAEP). These responses give some information about the integrity of the auditory nerves and auditory nuclei in the pons. Middle Latency Evoked Potential (MLEP) reflects activation of the primary auditory cortex. We can test higher function of the auditory cortex by increasing the complexity of the stimulus. With an "odd-ball" paradigm, we can record mismatch negativity (MMN). Its presence reflects some kind of pre-attentional auditory memory. Stimuli can be simple clicks or more complicated sounds (e.g., names).

Somato-sensory pathway: this pathway is tested by using electrical transcutaneous stimulation of peripheral nerves. Peripheral, medullar and cortical integrity can be tested. The cortical zone expected to be active is confined to the post-rolandic region (SSEP).

Motor pathway, Motor Evoked Potential (MEP) can be recorded with EMG surface electrodes after pre-rolandic stimulation, usually with Transcranial Magnetic Stimulation, but electric stimulation is also possible.

Several of the standard ERPs tests have been already applied to VS patients (Boris Kotchoubey et al. 2002) to assess sensory and cognitive functions. They have concluded that one- to two-thirds of patients with suspected VS are capable of cortical differentiation of physical stimulus features and that at least 20 % of these patients are able to differentiate semantic stimuli (i.e., their brains comprehend language). Interestingly, no difference seems to exist—according to these tests that rely on neural signals—between the minimally responsive and nonresponsive patients in language understanding. Indeed, at least three "nonresponsive" patients did differentiate words according to their semantic content. The authors (B. Kotchoubey and Lang 2001) attribute this result to the continuity of borders between typical and atypical VS and the difficulty of clinical differentiation between a "lack of responses" versus "weak and inconsistent responses."

Aforementioned tests are very well suited for the first loop of the PMI-TT as they probe the intactness of the sensory modalities and the cognitive level. Most ERP's components are well typified and their validity corroborated by many years of experience in research and in clinics. However, they are not specifically designed to test the main components of the definition of consciousness related to awareness of ourselves and our environment. Also, the ERP technique relies on the average of many repetitions of similar stimuli and therefore assumes that responses can be reliably evoked over long periods. However, we know that brain damaged patients show considerable fluctuations in alertness which should be reflected in neural signals and that we need responses to every single question if we want to engage in a reliable dialogue with the patient. This is why, and in contrast to previous work done in the EEG domain, we would like to use our experience in the recognition of patterns recorded during single responses to create a simple yes or no device for patients.

Despite considerable discussions about the definition of consciousness, most researchers agree that: (1) self-awareness and (2) awareness of the environment are components of it. Therefore more specific tests need to be devised to probe both components which are at the same time sufficiently specific to distinguishing,

at the single response level, nonconscious reflexes and instinctual responses from conscious responses.

1. *Testing for awareness of the self:* Awareness of the self and memory are intimately related (Gallup 1970). During the proposed PMI-TT, patients will require short-term memory capabilities to keep in mind the commands given by the examiner. On the other hand self-awareness completely depends on episodic (autobiographical) memory. Indeed, several pieces of evidence have been found showing that success in the mirror test (Gallup 1970), a classical test for consciousness, depends on the existence of memory of the self and of the environment (Howe and Courage 1997).
   Our battery of questions will assess the intactness of autobiographical memories on two different grounds: (a) Through the presentation of questions requiring the understanding of language and (b) Through the presentation of songs (music but not lyric) known to belong to the patient's own experience. Questions designed to test autobiographical memories should be carefully formulated to obtain robust neural (EEG) responses able to differentiate between positive and negative answers. Robustness is enhanced by averaging multiple responses to the same (or similar) questions which enhances the signal to noise ratio. This procedure requires a tight EEG alignment by the onset of the question itself and therefore precise triggers need to be inserted on the EEG data. Therefore, the right question to pose to the patients to exploit their neural signals is not "what is your name." Actually, the question must be stated in the form of an assertion—Your name is . . .. . .—followed by the correct or false name. A very small delay between the word "is" and the proposed name allows inserting a trigger on the EEG signal for the required averaging. Incorrect assertions in conscious and non-amnesic patients should be detected by the brain as a conflict and generate outcome related signals which are absent for correct responses. As a consequence outcome signals could permit the extraction of conclusions on the capacity of the patient to recognize himself. For the second type of question we will present to patients with classical melodies known to him (e.g., from infancy) but on some of the trials the melody will be digitally altered. Differential responses should appear between true and altered melodies to decide that a patient passed the test.
   While succeeding in previous tests for awareness of the self allows making inference about residual consciousness, failing in these tests, does not mean that patients are not conscious. Actually, patients recovering from coma after traumatic brain injury are known to show deep troubles with remote memories (amnesia is common) even if they are clearly situated and conscious.
2. *Awareness of the self-environment in amnesic patients:* Responding—with neural signals—to the commands prompted by the examiner is obviously a way to pass the Turing test for awareness of the environment. There is however a simple test that can be implemented with the help of the experience we have accumulated on BCI. The test exploits the close relationship between awareness

and attention. While these two terms are far from identical, it is clear that we become aware of the things we attend to and not from those that our attention ignores. In addition, we want to exploit the major role played by feedback signals coming from the peripheral nervous system (efferent signals) in the awareness of our own body. For instance, patients with one anesthetized limb are very likely to hit objects with it indicating that blocking transmission of electric impulses from limbs is essential for the awareness of parts of our body. We believe that an amnesic but conscious patient should be able to posit his/her attention on specific parts of the body upon command. However, the place where the patient is placing his/her attention when prompted can be easily determined in the complete absence of overt motor responses by exploiting this afferent information and the concept of steady state responses. Steady state responses are oscillatory brain patterns evoked by rapid and repetitive stimulus sequences and which differ from transient evoked responses to single stimuli presented briefly. The resulting oscillating response to the repetitive stimuli can be recorded in EEG. Steady state responses have been described in nearly all sensory modalities and are maximal for electrodes covering the sensory cortex linked to the stimulated modality.

Steady state responses are observed at all recording levels, i.e., from single neurons to scalp EEG and are strongly modulated by attention. Indeed, we have been using for several years now the SS responses in the visual modality for the accurate and fast control of a BCI. However, if the somatosensory pathway is relatively intact in patients we will rely on Somatosensory steady-state potentials (SSSEPs) as they only require attention and awareness to be evoked. SSSEPs can be evoked by amplitude-modulated mechanical (vibrotactile) stimuli or currents (electrical). Stimuli are typically applied to the fingers and/or palmar surface of the hand. Response's amplitudes are greatest at low frequencies. The greatest signal-to-noise ratio is found at modulation frequencies near 26 Hz and response latency is about 58 ms. To test residual consciousness and eventually obtain a binary communication window with the patient, we plan to simultaneously stimulate his left and right hand with tiny currents or vibrotactile stimuli of different frequencies. We will then ask the patient to concentrate on what is going on in his/her hands. If he is able to understand the command and place the attention on the prompted hand we should be able to detect clear increases in the power of the frequency that is given at that hand and no changes in the other. Passing this test will indicate that: (1) a patient is able to understand spoken language, (2) he is aware of his environment as he respond to our prompt and (3) he is aware of his own body and attentive.

To study the sensory and cognitive functions in the VS patients we will implement previous tests using free software for cognitive stimuli presentation. Both, ERP responses and multivariate pattern recognition approaches will both help us to clarify neural functions in individual patients.

10.4 Pattern Recognition for the Assessment of Consciousness and to Establish a Basic Dialog with Patients

Statistical pattern recognition algorithms are designed to learn and later classify multivariate data points based on statistical regularities in the data set. Learning is based on selecting some patterns (features) over one part of the trials (the learning set created during the initial learning session). We then give these patterns to the classifier along with a label that identifies the frequency of the stimuli that subjects were instructed to attend. The classifier learns a mapping between patterns of brain activity (the neural oscillations) and the presented stimuli. Using diverse statistical measures we can rank the features and build a linear classifier based on the best features and the Proximal SVM approach (PSVM). Finally a heuristic filtering strategy is added to the output of the classifier to suppress false positives. Note that the term classifier here refers to both the PSVM and the filtering strategy of the output scores. Consequently, the PMI-TT proposed here is based on very fast algorithms allowing an efficient online implementation in case a dialogue could be established with the patient.

10.5 Conclusions

Our ultimate goal is to implement a complete battery of tests that allows the quantification of consciousness within clinical settings in a more precise way. The main difficulty is that consciousness is not uniquely defined and we don't therefore know the conditions that are necessary or sufficient to assert that patients are conscious. We believe that the operational definition presented here based on the awareness of the self/environment might however help to distinguish between minimally conscious and completely unconscious patients. Hence, this proposal adopts the "soft" view on the problem of consciousness that is compatible with the physical idea that traces of a phenomenon can be measured even if a full understanding of the phenomenon is lacking.

This proposal leads immediately to the question of whether or which patients retain sufficient consciousness as to imagine that BCIs might improve their welfare and become their way to communicate with the world. A compelling evidence to ascertain that VS patients are conscious is to prove that they manage to understand verbal instructions and comply with some requests. Focally and selectively engaging attention in parts of their body and ignoring others when specifically instructed is a way to satisfactorily pass the operational test for consciousness which is both, simple to apply and independent on the remaining long term memory capacity of patients. This could therefore constitute the basis to eventually establish a dialogue and therefore become their only way to communicate with the world.

It is clear that the clinical goals aimed here cannot provide a definitive answer to the problem of what is consciousness but rather help to pinpoint what it is not.

The rationale undertaken is that if consciousness exists we should be able to measure it. Reverse neural engineering, as today understood in neuroscience, could help within this framework to ultimately lead to a better understanding of consciousness. As previously argued consciousness is not memory (as an amnesic patient can be totally conscious) but apparently depends on it. In the same vein, consciousness is not identical to attention but the capacity to become aware of things correlates with our capacity to attend to them. Consequently, it is hard to believe that consciousness could be reduced to the activation of few brain areas. It is easier to think in consciousness as a highly dynamic and emergent property of complex systems. Under this view the study of conscious processes by means of the highly dynamic but eye bird view provided by the scalp EEG seems fully adequate. We will complement scalp EEG with electrical neuroimaging, i.e., the determination of neural sources from scalp EEG via the solution of the inverse problem (Grave de Peralta Menendez et al. 2000; Rolando Grave de Peralta Menendez et al. 2004). We hope that by combining reverse engineering in patients with electrical neuroimaging we will be able to shed further light on the soft problem of consciousness. The advantages of the combination were already illustrated in the phenomenon known as blindsight (Gonzalez Andino et al. 2009) related to consciousness in visual perception. The hard problem of consciousness—how it emerges—is certainly much more difficult to tackle but progresses in any direction are likely to help.

References

Andrews, K., Murphy, L., Munday, R., & Littlewood, C. (1996). Misdiagnosis of the vegetative state: Retrospective study in a rehabilitation unit. *British Medical Journal, 313*(7048), 13–16.

Childs, N. L., Mercer, W. N., & Childs, H. W. (1993). Accuracy of diagnosis of persistent vegetative state. *Neurology, 43*(8), 1465–1467.

Gallup, G. G. (1970). Chimpanzees: Self recognition. *Science, 167*(914), 86–87.

Gonzalez Andino, S. L., de Peralta, G., Menendez, R., Khateb, A., Landis, T., & Pegna, A. J. (2009). Electrophysiological correlates of affective blindsight. *Neuroimage, 44*(2), 581–589.

de Peralta, G., Menendez, R., Gonzalez Andino, S. L., Morand, S., Michel, C. M., & Landis, T. (2000). Imaging the electrical activity of the brain: ELECTRA. *Human Brain Mapping, 9*(1), 1–12.

de Peralta, G., Menendez, R., Murray, M. M., Michel, C. M., Martuzzi, R., & Gonzalez Andino, S. L. (2004). Electrical neuroimaging based on biophysical constraints. *Neuroimage, 21*(2), 527–539.

Howe, M. L., & Courage, M. L. (1997). The emergence and early development of autobiographical memory. *Psychological Review, 104*(3), 499–523.

Kotchoubey, B., & Lang, S. (2001). Event-related potentials in an auditory semantic oddball task in humans. *Neuroscience Letters, 310*(2–3), 93–96. http://www.sciencedirect.com/science/article/pii/S0304394001020572

Kotchoubey, B., Lang, S., Bostanov, V., & Birbaumer, N. (2002). Is there a mind? Electrophysiology of unconscious patients. *News in Physiological Sciences, 17*(1), 38–42.

Laureys, S., Owen, A. M., & Schiff, N. D. (2004). Brain function in coma, vegetative state, and related disorders. *Lancet Neurology, 3*(9), 537–546. doi:10.1016/S1474-4422(04)00852-X.

Overgaard, M. (2009). How can we know if patients in coma, vegetative state or minimally conscious state are conscious? Progress in brain research. In N. D. S. a. A. M. O. Steven Laureys (Ed.), *Coma science: Clinical and ethical implications* (Vol. 177, pp. 11–19). Elsevier.

Owen, A. M., Schiff, N. D., & Laureys, S. (2009). The assessment of conscious awareness in the vegetative state. In S. Laureys & G. Tononi (Eds.), *The neurology of consciousness* (pp. 163–172). San Diego: Academic.

Regan, D. (1989). *Human brain electrophysiology: evoked potentials and evoked magnetic fields in science and medicine*. New York: Elsevier.

Schneider, S., & Velmans, M. (2007). Introduction to the Blackwell companion to consciousness. In M. V. S. Schneider (Ed.), *The blackwell companion to consciousness* (pp. 1–6). Blackwell.

Chapter 11
On the Irreducibility of Consciousness and Its Relevance to Free Will

Giulio Tononi

Abstract Integrated information theory of consciousness (IIT) starts from phenomenological axioms and argues that an experience is an *integrated information structure*. IIT holds that a system of connected elements—for example a network of neurons, some firing and some not—intrinsically and necessarily generates information, because its mechanisms and present state constrain possible past and future states. This intrinsic, causal kind of information—called *cause-effect information* (CEI)—measures "differences that make a difference" from the intrinsic perspective of the system. Moreover, a subset of elements only generates information to the extent that the cause and effect repertoires they specify cannot be reduced to the product of the repertoires specified by independent components (*integrated information*, φ). Finally, only *maxima of integrated information* ($_{\max}\varphi$) matter. A maximally irreducible cause-effect repertoire constitutes a *concept*. A *complex* is a set of elements specifying a maximally irreducible constellation of concepts ($_{\max}\Phi$), giving rise to a maximally integrated conceptual information structure or *quale*. Under certain conditions, such as the presence of noise and irreversibility, a maximum of integrated information may be associated with a "macro" spatiotemporal grain (say neurons over hundreds of milliseconds), rather than with a "micro" grain (say subatomic particles over microseconds). IIT accounts, in a parsimonious manner, for many, seemingly disparate empirical observations about consciousness, and makes theoretical predictions concerning the necessary and sufficient conditions for the presence and quality of consciousness in newborns, brain damaged patients, animals, and machines. Moreover, IIT has direct relevance for issues related to free will. According to IIT, when a choice is made consciously, in addition to satisfying the requirements of autonomy, understanding, self-control, and alternative possibilities, the choice is maximally irreducible. This is because the choice cannot be attributed to anything less than the entire complex that brings it about, nor is

G. Tononi (✉)
Department of Psychiatry, University of Wisconsin, 6001 Research Park Boulevard, Madison, WI 53719, USA
e-mail: gtononi@wisc.edu

A. Suarez and P. Adams (eds.), *Is Science Compatible with Free Will?: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience*,
DOI 10.1007/978-1-4614-5212-6_11,

anything more than the complex required, as the complex provides the maximally irreducible set of cause-effects. If maximal integrated information is generated by a complex at a macroscale in space or time (groups of neurons, hundreds of milliseconds), the requirement for indeterminism is also satisfied: a conscious choice, while maximally and irreducibly causal, is also necessarily under-determined and thus unpredictable. In this view, indeterminism is not to be thought of as a sprinkle of randomness that instills some arbitrariness into a preordained cascade of mechanisms, decreasing their causal powers. Rather, indeterminism provides a backdrop of ultimate unpredictability against which information integration acts to impose autonomy, understanding, self-control, and alternative possibilities. Thus, according to IIT, a choice is the freer, the more it is determined intrinsically, meaning that it can only be accounted for by considering a large set of concepts, beliefs, memories, and wishes, all acting within a maximally irreducible complex. Which is to say that a choice is the freer, the more it is conscious.

Keywords Information • Integration • Perception • Action • Causality • Reductionism

11.1 Axioms and Postulates of Consciousness as Integrated Information

The main tenets of integrated information theory (IIT) of consciousness can be presented as a set of phenomenological axioms, ontological postulates, and identities. The central axioms, which are taken to be immediately evident, are as follows:

An initial axiom is simply that *consciousness exists*. Paraphrasing Descartes, "*I experience therefore I am*."[1]

Another axiom concerns compositionality: *experience is structured, consisting of multiple aspects in various combinations*. Thus, even an experience of pure darkness and silence contains visual and auditory aspects, spatial aspects such as left center and right, and so on.

A central axiom concerns information: *experience is informative* or *specific*—in that *it differs in its particular way from other possible experiences*. Thus, an experience of pure darkness and silence is what it is by differing, in its particular way, from an immense number of other possible experiences—including the experiences triggered by any frame of any possible movie.

Another axiom concerns integration: *experience is integrated*—in that *it cannot be reduced to independent components*. Thus, experiencing the word "SONO" written in the middle of a blank page cannot be reduced to an experience of the word "SO" at the right border of one page, plus an experience of the word "NO" on the left border of another page—the experience is whole.

[1] Descartes started his philosophical investigations from the axiom "cogito ergo sum," though his "cogito" emphasized the thinking aspect of consciousness rather than the more general notion of having an experience.

Yet another axiom is exclusion: *experience is exclusive*—in that *it has definite borders, temporal, and spatial grain*. Thus, an experience encompasses what it does, and nothing more; it flows at a particular speed, and it has a certain resolution such that certain distinctions are possible and finer distinctions are not.

To parallel the phenomenological axioms, IIT posits some *ontological postulates*:

An initial postulate is simply that *mechanisms in a state exist*. That is, there are operators that, given an input, produce an output, and at a given time such operators are in a particular state.

Another postulate concerns *compositionality: mechanisms can be structured, forming higher order mechanisms in various combinations*.

A central postulate concerns *information*: *from the intrinsic perspective of a system, a mechanism in a state generates information only if it has* both *specific causes* and *specific effects* within *the system*—that is, the mechanism must constitute "a difference that makes a difference within the system." This intrinsic, causal notion of information can be assessed by partitioning the mechanism from the system on the input *or* output side: if this reduction to just inputs or just outputs makes no difference to the system, then the mechanism does not generate any cause-effect information (CEI) within the system. Ontologically, the information postulate claims that, from the intrinsic perspective of a system, only differences that make a difference within the system exist.

Another postulate concerns *integration*: *a mechanism in a state generates integrated information only if it cannot be partitioned into independent submechanisms*. That is, the information generated within a system *should be irreducible* to the information generated within independent subsystems or independent interactions. Integrated information (φ) can be captured by measuring to what extent the information generated by the whole differs from the information generated by its partitioned components. Ontologically, the integration postulate claims that only irreducible interactions exist intrinsically, i.e., in and of themselves.

Yet another postulate concerns *exclusion*: *a mechanism in a state generates integrated information about only one subset of causes and effects—the one that is maximally irreducible*. That is, the mechanism can specify only one pair of causes and effects. By a principle of causal parsimony, this is the pair of causes and effects whose partition would produce the greatest loss of information. This maximally irreducible set of causes and effects is called a *concept*. Exclusion can be captured by measuring the maximum of integrated information $_{\max}\varphi$ over all possible cause-effect repertoires (CERs) of the mechanism over the system. Ontologically, the exclusion postulate claims that only maximally irreducible entities exist intrinsically.[2]

As will be discussed below, the postulates can be applied to subsets of elements within a system (mechanisms) as well as to systems (sets of concepts). A system of elements that generates cause-effect information (it has concepts), is irreducible

[2] Contrasting with this *intrinsic* perspective, which is observer-independent, is the *extrinsic* perspective of an external observer: the observer can ask how information is encoded, communicated or stored given the system's state and the observer's expectations (prior distribution, e.g., based on observing the system) and assumptions about the system.

(it cannot be split into independent subsystems), and is a local maximum of irreducibility (in terms of the concepts it generates) over an *optimal spatio-temporal grain* of interactions, constitutes a *complex*—a maximally irreducible entity. In this view, only complexes are entities that exist intrinsically, i.e., in and of themselves.

Finally, IIT posits *identities* between phenomenological aspects and informational/causal aspects of systems. The central identity is the following: *an experience is a maximally integrated information structure*. Said otherwise, an experience is *a* "shape" *in qualia space* (a *quale*), where qualia space is a space spanned by all possible past and future states of a complex. In this space, concepts are points in the space whose coordinates are the probabilities of past and future system states corresponding to maximally irreducible CERs specified by various subsets of elements.

In what follows, the postulates of IIT are briefly illustrated by considering how they can be applied to individual mechanisms in a state (concepts), and then to a collection of mechanisms (complexes).

11.2 Information

The *information* postulate says that *information is a difference that makes a difference from the intrinsic perspective of a system.* This intrinsic, causal notion of information is assessed by considering if the present state of a mechanism can specify both past causes and future effects within the system.

Within a system X, consider a subset of elements S in its present state s. The information s generates about some subset of elements of X in the past (P) is the *effective information* (EI) between P and s:

$$\mathrm{EI}(P|s) = (P|s)||P^{\mathrm{Hmax}}$$

where || indicates the distance D[3] between two distributions, in this case between the distribution of P states that could have caused s given its present mechanism and state (the *cause repertoire* CR), and the maximum uncertainty (entropy) distribution P^{Hmax}, in which all P outputs are equally likely a priori. Thus, EI($P|s$) represents the differences in the past states of P that that can be detected by mechanism S in its present state s. Similarly, the distance D between the distribution of F states that would be the effect of "fixing" mechanism S in its present state s (the *effect repertoire* ER) and the distribution of states of F in which all F inputs are equally likely (F^{Hmax}), is the effective information s generates about future states of F:

$$\mathrm{EI}(F|s) = (F|s)||F^{\mathrm{Hmax}}$$

[3] The distance D between two probability distributions p and q can be measured in various ways. Perhaps the most general way is to consider the information distance between them, i.e. the maximum of the Kolmogorov complexity of one distribution given the other (Bennett et al. 1998). See Tononi (2013) for further considerations.

Thus, EI($F|s$) represents the differences to the future states of F made by mechanism S being in its present state s. Clearly, EI($P|s$) > 0 only if past states of P make a difference to s, and EI($F|s$) > 0 only if s makes a difference to F. Based on the information postulate, a mechanism in a state (s) generates information from the intrinsic perspective of a system only if it *both* detects differences in the past states of the system *and* makes a difference to its future states. That is, s generates information only if it has *both* specific causes (EI($P|s$) > 0) *and* specific effects (EI($F|s$) > 0). The minimum of the two, which represents the "bottleneck" in the channel between past causes over P and future effects over F as mediated by the mechanism S in its present state s, is called *cause-effect information* (CEI):

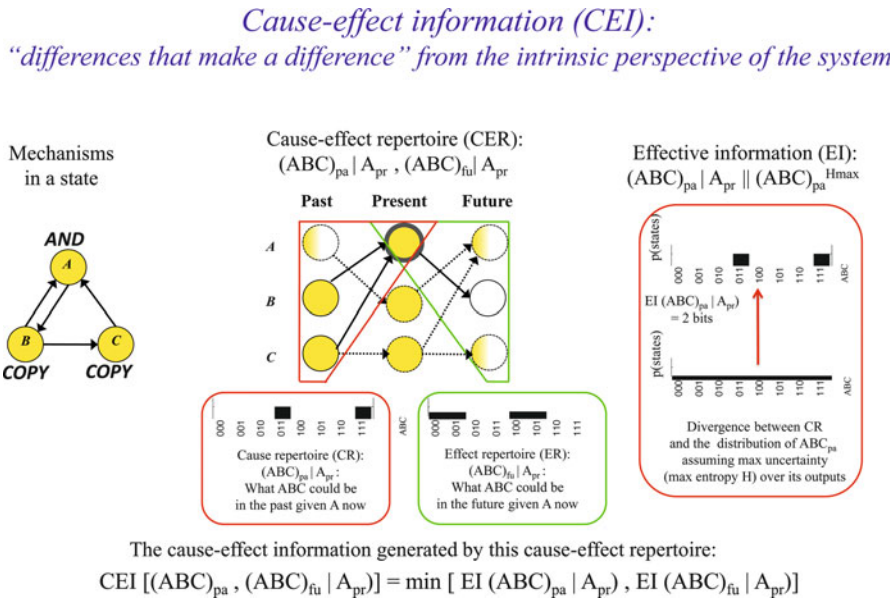
$$\mathrm{CEI}(P, F|s) = \min[\mathrm{EI}(P|s), \mathrm{EI}(F|s)]$$

Clearly, CEI > 0 only if the system's states make a difference to the mechanism, *and* the state of the mechanism makes a difference to the system. Thus an element that monitors the state of the system (say a parity detector), but has no effects on the system, may be relevant from the extrinsic perspective of an observer, but is irrelevant from the intrinsic perspective of the system, as it makes no difference to it. If CEI > 0, the cause and effect repertoires together can be said to specify a CER.

As an example, consider a mechanism A within an isolated system ABC (Fig. 11.1). The wiring diagram is unfolded into a directed acyclic graph over past, present, and future. A's mechanism is a logical AND gate of elements B and C, turning ON if both B and C are ON; moreover, if A is ON, it turns OFF B. Thus, A specifies that, starting from the eight possible past states of elements ABC (maximum entropy distribution), only two past outputs of ABC can lead to A's present state (ON)—those in which B and C are both ON (cause repertoire CR), thereby "detecting differences" and generating EI. Moreover, A specifies that, starting from maximum entropy over the inputs to ABC, A's present state (ON) can only lead to four future states of ABC—those in which B is OFF (effect repertoire ER), thereby "making a difference." Together, CR and ER specify the cause-effect repertoire CER $= (\mathrm{ABC})_{\mathrm{pa}} \mid \mathrm{A}_{\mathrm{pr}}, (\mathrm{ABC})_{\mathrm{fu}} \mid \mathrm{A}_{\mathrm{pr}}$, where the subscripts refer to present, past, and future. The cause-effect information (CEI) generated by a mechanism over its CER is the minimum between EI $[(\mathrm{ABC})_{\mathrm{pa}} \mid \mathrm{A}_{\mathrm{pr}}]$ and EI $[(\mathrm{ABC})_{\mathrm{fu}} \mid \mathrm{A}_{\mathrm{pr}}]$.

11.3 Integration

The *integration* postulate says that *information is integrated if it cannot be partitioned into independent components*. That is, *a mechanism in state generates integrated information only if it cannot be partitioned into submechanisms with independent causes and effects*. This integrated (irreducible) information is quantified by φ (small phi), a measure of the distance D between the repertoire specified by a whole and the product of the repertoires specified by its partition into



$$\text{CEI}\,[(\text{ABC})_{pa}, (\text{ABC})_{fu} \mid \text{A}_{pr})] = \min\,[\,\text{EI}\,(\text{ABC})_{pa} \mid \text{A}_{pr}),\ \text{EI}\,(\text{ABC})_{fu} \mid \text{A}_{pr})]$$

Fig. 11.1 A cause-effect repertoire (CER) and the cause-effect information it generates ("differences that make a difference"). See text for explanation

causally independent components. The distance is taken over the partition that yields the least distance from the whole (the *minimum information partition* (MIP)), i.e., φ^{MIP}.[4]

Consider a partition **x** that splits the interactions between P and S into independent interactions between parts of P and parts of S,[5] which can be done by "injecting" noise ($H^{\max}$) in the connections among them. One can then measure the distance D between the unpartitioned cause repertoire CR and the partitioned CR. For the partition that minimizes D, known as *minimum information partition* (MIP), the distance is called φ (small phi). The same holds for the distance between the unpartitioned and partitioned effect repertoire ER:

$$\varphi^{\text{MIP}}(P|s) = (P|s)||\prod(P|s/\text{MIP}); \quad \varphi^{\text{MIP}}(F|s) = (F|s)||\prod(F|s/\text{MIP})$$

Thus, $\varphi^{\text{MIP}}(P|s)$ is the "past" *integrated* (*irreducible*) *information*, and $\varphi^{\text{MIP}}(F|s)$ is the "future" *integrated* (*irreducible*) *information*. Clearly, $\varphi^{\text{MIP}}(P|s) > 0$ only if the past states of P make a difference to s that cannot be reduced to differences made by parts of P on parts of s, and likewise for $\varphi^{\text{MIP}}(F|s) > 0$.

[4] Partitions, indicated by **x,** can be evaluated by performing the same computations after injecting noise (do($H^{\max}$)) in the partitioned links in the input–output matrix. To fairly compare different partitions to find the MIP, it is necessary to normalize by the information capacity of each partition.

[5] Where the empty set [] is only allowed on either P or S, but not both.

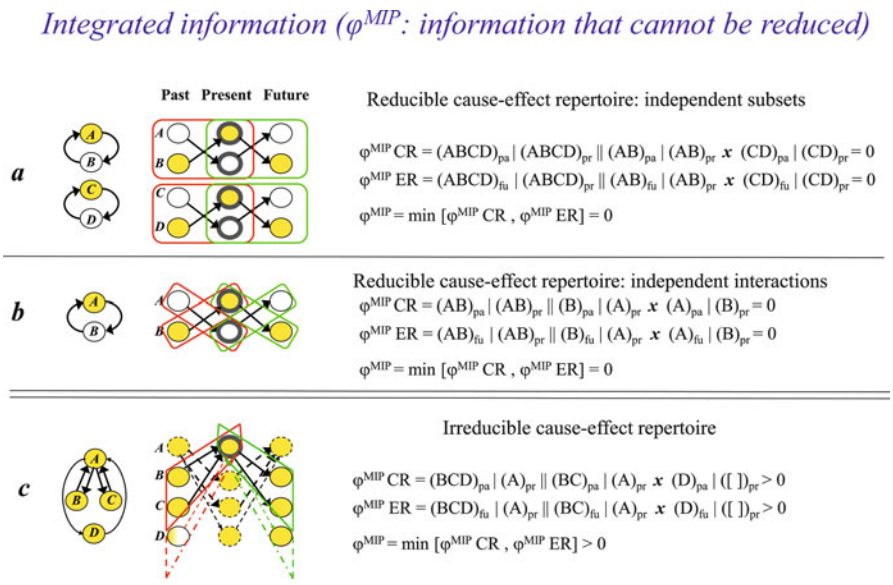


Fig. 11.2 Integrated information generated by an irreducible CER, as established by performing partitions. See text for explanation

Based again on the information postulate, a mechanism in a state (s) generates integrated information from the intrinsic perspective of a system only if this information is irreducible *both* in the past *and* in the future. That is, s generates integrated information only if it has *both* irreducible causes ($\varphi^{\mathrm{MIP}}(P|s) > 0$) *and* irreducible effects ($\varphi^{\mathrm{MIP}}(F|s) > 0$). The minimum of the two, which represents the "bottleneck" in the channel between the past P and the future F as mediated by the mechanism S in its present state s, is called "cause-effect" *integrated information:*

$$\varphi^{\mathrm{MIP}}(P, F|s) = \min[\varphi^{\mathrm{MIP}}(P|s), \varphi^{\mathrm{MIP}}(F|s)]$$

As an example, Fig. 11.2a shows a set of four elements ABCD, where A is reciprocally connected to B and C is reciprocally connected to D. The wiring diagram is again unfolded into a directed acyclic graph over past, present, and future. Consider now the cause repertoire $(ABCD)_{pa} \mid (ABCD)_{pr}$ and a partition between subsets of elements AB on one side and CD on the other side: $\varphi^{\mathrm{MIP}}(P \mid s) = (ABCD)_{pa} \mid (ABCD)_{pr} \parallel (AB)_{pa} \mid (AB)_{pr}$ **x** $(CD)_{pa} \mid (CD)_{pr} = 0$. Similarly for the effect repertoire, $\varphi^{\mathrm{MIP}}(F \mid s) = (ABCD)_{fu} \mid (ABCD)_{pr} \parallel (AB)_{fu} \mid (AB)_{pr}$ **x** $(CD)_{fu} \mid (CD)_{pr} = 0$. Thus, as expected, for this partition $\varphi^{\mathrm{MIP}} = \min [\varphi^{\mathrm{MIP}}(P \mid s), \varphi^{\mathrm{MIP}}(P \mid s)] = 0$. That is, considering the "whole" CER specified by $(ABCD)_{pa} \mid (ABCD)_{pr}$ and $(ABCD)_{fu} \mid (ABCD)_{pr}$ adds nothing compared to considering the independent "partial" CER specified by $(AB)_{pa} \mid (AB)_{pr}$, $(AB)_{fu} \mid (AB)_{pr}$ and by $(CD)_{pa} \mid (CD)_{pr}$, $(CD)_{fu} \mid (CD)_{pr}$. In other words, there is no reason to maintain that the "whole" CER ABCD exists in and of itself, as it makes no difference above and beyond the two partial CER AB and CD. Thus, searching for partitions among sets of elements yielding $\varphi^{\mathrm{MIP}} = 0$ enforces a principle of causal parsimony.

As another example, consider a partition between interactions. The system depicted in Fig. 11.2b is such that A copies B and B copies A. For the cause-repertoire CR of AB and its partition into independent interactions of A with B and B with A one has that φ^{MIP} $(P \mid s) = (\mathrm{AB})_{\mathrm{pa}} \mid (\mathrm{AB})_{\mathrm{pr}} \parallel (\mathrm{B})_{\mathrm{pa}} \mid (\mathrm{A})_{\mathrm{pr}}$ **x** $(\mathrm{A})_{\mathrm{pa}} \mid (\mathrm{B})_{\mathrm{pr}} = 0$, and similarly for the effect repertoire ER. That is, the CER of AB over AB (written AB/AB) reduces without loss to the independent CER of A/B and B/A both in the past and in the future. Thus, there is no reason to maintain that the CER AB/AB exists in and of itself, as it makes no difference above and beyond the independent CER of A/B and B/A. Again, searching for partitions among interactions yielding $\varphi^{\mathrm{MIP}} = 0$ enforces a principle of causal parsimony.

By contrast, consider a system in which A is a linear threshold unit that receives strong inputs from B and C, which if both ON are sufficient to turn A ON, and a weak input from D; and in which A has strong outputs to B and C (it turns both ON), and a weak output to D (Fig. 11.2c). Considering the CR of A/BCD, one has that its partition A/BC **x** D/[] ([] indicates the empty set) yields $\varphi^{\mathrm{MIP}} > 0$, and the same holds for the ER. Thus, this CER is irreducible, since there is no way to partition it without losing some information—in this case some information about element D.

11.4 Exclusion

The *exclusion* postulate says that *integrated information is about only one subset of causes and effects—those that are maximally irreducible*. That is, a mechanism in a state can specify only one pair of causes and effects, which, by a principle of causal parsimony, is the one whose partition would produce the greatest loss of information. This *maximally irreducible set of causes and effects* (MICE) is called a *concept* or, for emphasis, a "core concept."

For a given subset of elements S in a present state s, there are potentially many cause repertoires CR depending on the particular subset P one considers (within system X). Exclusion states that, at a given time, s can have only one CR—which is the one having the maximum value of φ^{MIP} ($_{\max}\varphi^{\mathrm{MIP}}$), where the maximum is taken over all possible subsets P within the system.[6] The corresponding CR is called the "core" *cause* of s within X. Similarly, the effect repertoire ER having $_{\max}\varphi^{\mathrm{MIP}}$ over all possible subsets F within the system is called the "*core*" *effect* of s within X.

Based again on the information postulate, a mechanism in a state (s) generates maximally integrated information from the intrinsic perspective of a system only if this information is maximally irreducible *both* in the past *and* in the future. That is, s generates maximally integrated information only if it has *both* maximally

[6] If several CER(S) yield the same max, one takes the CER(S) of largest scope (accounting for the most), where $\varphi^{\mathrm{MIP}}(S) > 0$, its subsets R have lower or at most equal φ^{MIP}, and its supersets T have lower φ^{MIP}: φ^{MIP} $(R) \leq \varphi^{\mathrm{MIP}}$ $(S) > \varphi^{\mathrm{MIP}}$ (T), for all $R \subset S$ and all $T \supset S$. If there are multiple maximal CER(S) each with the same scope, then at any given time only one is realized as a concept, although which one is indeterminate.

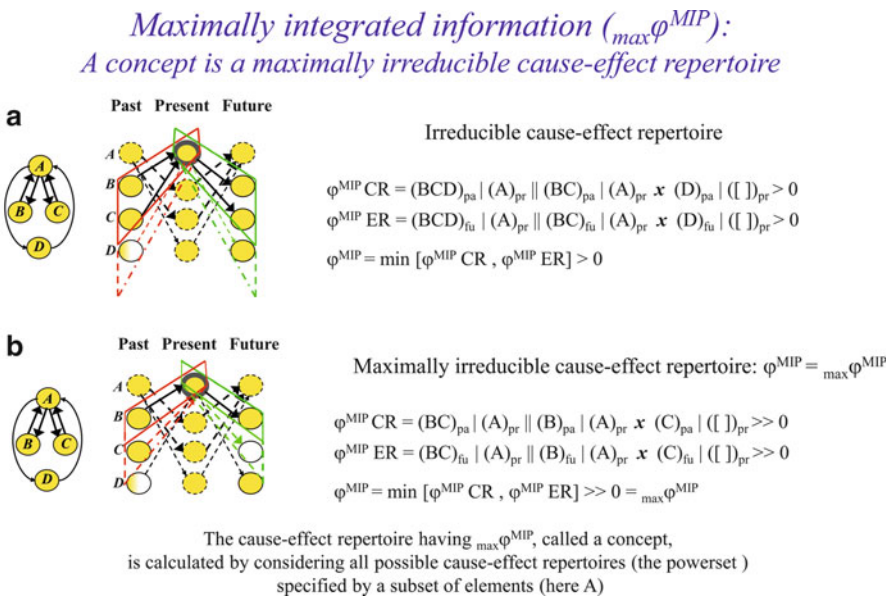


Fig. 11.3 Maximally integrated information generated by a maximally irreducible CER over all possible CER specified by a subset of elements within a system. See text for explanation

irreducible causes $({}_{\max}\varphi^{\mathrm{MIP}}(P|s) > 0)$ *and* maximally irreducible effects $({}_{\max}\varphi^{\mathrm{MIP}}(F|s) > 0)$. The minimum of the two, which represents the "bottleneck" in the channel between the past P and the future F as mediated by the mechanism S in its present state s, is called "cause-effect" *maximally integrated information:*

$$ {}_{\max}\varphi^{\mathrm{MIP}}(P, F|s) = \min[{}_{\max}\varphi^{\mathrm{MIP}}(P|s),\ {}_{\max}\varphi^{\mathrm{MIP}}(F|s)] $$

The CER of s that has $_{\max}\varphi^{\mathrm{MIP}}$ $(P, F|s)$ within a system X is called a *concept*. Thus, from the intrinsic perspective of a system, a concept is a *maximally irreducible set of causes and effects* (MICE) specified by a mechanism in a state.

For example, in Fig. 11.3 the powerset of CER of subset A within system ABCD includes, for the cause repertoires, A/A; A/B; A/C; A/D; A/AB; A/AC; A/AD; A/BC; A/BD; A/CD; A/ABC; A/ABD; A/ACD; A/BCD; A/ABCD. Of these, the partition A/BC ∥ A/B **x** []/C = $_{\max}\varphi^{\mathrm{MIP}}$ turns out to be maximal (Fig. 11.3b), higher for example than the partition A/BCD ∥ A/BC **x** []/D in Fig. 11.3a. This is because partitioning away element B (or A) loses much more integrated information than any other partition. A similar result is obtained for the powerset of partitions of A/ABCD for the effect repertoires. By the exclusion postulate, only one CER exists—the one made of the maximally irreducible CR and ER—excluding any other CER.

The reason to consider exclusively the CER with $_{\max}\varphi^{\mathrm{MIP}}$ is as before a principle of causal parsimony—more precisely, a *principle of least reducible reason*. Consider A being ON in the previous example: it specifies a cause repertoire, but cannot distinguish which particular cause was actually responsible for its being ON;

and with respect to its effects, it makes no difference which cause turned A ON. Since the particular cause does not matter, the exclusion postulate enforces causal parsimony, defaulting to the maximally irreducible set of causes for A being ON. These least "dispensable" and thus most likely "responsible" causes can be called the "core" causes for A being ON, in the sense that their elimination would have made the most difference from the intrinsic perspective of A.[7,8] In turn, the fact that A is ON also specifies a forward repertoire of possible effects, but once again A should be held most responsible only for its maximally irreducible or "core" effects: the effects for A being ON is least dispensable, meaning that eliminating A's output would have made the most difference.[9]

11.5 Concepts

A concept or "core" concept thus specifies a maximally irreducible CER implemented by a mechanism in a state. Within a concept, one can distinguish a core *cause*—the set of past input states (cause repertoire CR) constituting maximally irreducible causes of the present state of the mechanism; and a core *effect*—the set of future output states (effect repertoire ER) constituting maximally irreducible effects of its present state. For example, an element (or set of elements) implementing the concept "table," when ON, specifies "backward" the maximally irreducible set of inputs that could have caused its turning ON (e.g., seeing, touching, imagining a table); "forward," it specifies the set of outputs that would be the effects of its turning ON (e.g., thinking of sitting at, writing over, pounding on a table).[10]

[7] One could say that trying various CER and their partitions to find $_{max}\varphi^{MIP}$ is the informational/causal equivalent of "cutting to the chase." It is also related to finding the optimal tradeoff between the transmission of relevant information and the compression/efficiency of the channel.

[8] In neural terms, the fact that, out of all possible causes of a neuron's firing, the input that actually caused its firing remains undecidable from the intrinsic perspective, also means that "illusions" are inevitable. Based on the exclusion postulate, the intrinsic perspective entails the simplifying attribution of cause always to the core (most irreducible) cause, rightly or wrongly. Usually, in an adapted system, the actual cause and the core cause will be similar enough, but occasionally the actual cause may be quite different from the core cause, in which case an "illusion" ensues (this applies also to the case of a neuron's firing being caused by microstimulation).

[9] The exclusion postulate is related to the principle of sufficient reason—in fact, it enforces a principle of *least* reducible reason; to the principle of least action; to maximum likelihood approaches and to information minimization/compression (though it is causal, not just statistical); and of course ultimately to Occam's razor.

[10] In this example, the cause repertoire component of a concept (backward, input, retrodictive, receptive concept) can be taken to refer to a classic *invariant*—a set of inputs equivalently compatible with the present state of a certain mechanism (e.g., tables, faces, places, and so on); the effect repertoire component (forward, output, predictive, projective concept) can be taken to refer to "Gibsonian" *affordances*—a set of outputs equivalently compatible with the present state of a certain mechanism (e.g., the consequences/associations/actions primed by seeing a table, face, place, and so on).

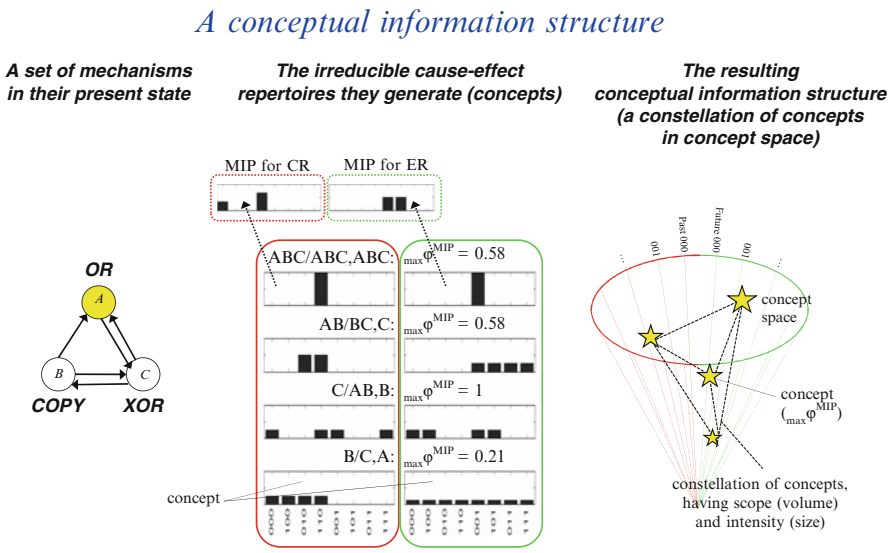


Fig. 11.4 An integrated conceptual information structure. See text for explanation

As an example, consider the system in Fig. 11.4, whose wiring diagram is on the left. The middle panel shows the four concepts generated by the system, with their maximally irreducible CERs and the corresponding $_{\max}\varphi^{\mathrm{MIP}}$. For the concept generated by all three elements (ABC, top row) the figure also shows the product repertoires generated by the minimum information partitions of its maximal cause and effect repertoires.

For a given set of elements, it is useful to consider concepts as points within a space (*concept space*) that has as many axes as the number of possible past and future states of the set (Fig. 11.4, right panel; the axes are depicted along a circle but should be imagined in a high-dimensional space; the points are indicated as stars). Each concept specifies a maximally irreducible CER, which is a set of probabilities over all possible past and future states, and these probabilities specify a particular point in concept space (more precisely, since probabilities must sum to 1, in the subspace given by the corresponding *concept simplex*). The concept "exists" with an "intensity" given by $_{\max}\varphi^{\mathrm{MIP}}$, that is, its degree of irreducibility (shown by the size of the star).

It is thus possible to evaluate the overall *constellation* of concepts generated by the set of elements in a single concept space, which can be called a *conceptual information structure*. Among the relevant features one can consider are: (1) the intensity, i.e., irreducibility $_{\max}\varphi^{\mathrm{MIP}}$ of existing concepts; (2) the constellation or "shape" in the space specified by the concepts; (3) the dimensionality of the subspace spanned by all the concepts; (4) the scope of the subspace covered by the concepts; (5) the scope of the subspace covered by the concepts weighted by their intensity.

11.6 Complexes

By considering the conceptual information structure specified in concept space by all the concepts a system generates (Fig. 11.4), the postulates of IIT can be applied not only to find the maximally irreducible CER of a subset of elements (concepts), but also to find sets of elements, called *complexes*, which generate maximally irreducible conceptual information structures.

As with concepts, so with complexes, this can be done by: i) making sure that a set of elements generates a constellation of concepts (information postulate); ii) evaluating to what extent the constellation of concepts generated by the set of elements is irreducible (integration postulate); iii) choosing the set of elements that generates the most irreducible constellation of concepts (exclusion postulate).

As before, the irreducibility mandated by the integration postulate can be determined by measuring the difference D between the constellation of concepts generated by the whole (unpartitioned set of elements s) and that generated by its minimal parts (partitioned set of elements s/P, where the partition is the minimum information partition (MIP)). The greater the distance D for the MIP, the more irreducible the constellation of concepts generated by a particular set of elements:

$$\Phi^{\mathrm{MIP}}(\mathrm{C}\,|\,\mathrm{s}) = \min\,[\,\mathrm{D}\,(\mathrm{C}\,|\,\mathrm{s},\ \mathrm{C}\,|\,\mathrm{s}/\mathrm{MIP}\,\rightarrow), \mathrm{D}(\mathrm{C}\,|\,\mathrm{s},\ \mathrm{C}\,|\,\mathrm{s}/\mathrm{MIP}\,\leftarrow)]$$

where Φ^{MIP} stands for *integrated conceptual information* (as for concepts, evaluating the distance D between the two constellations (here, the unpartitioned and the partitioned one) can be done by considering the information distance between them, i.e. the maximum of the Kolmogorov complexity of one given the other (Bennett et al. 1998); the arrow after MIP indicates that one should first partition across the inputs to one side of the partition (i.e. the outputs from the other side), then across the outputs (inputs), and take the minimum of the distance D made by the two one-directional cuts. The MIP is the partition for which this (normalized) value of this minimum is the least over all partitions. Thus, for integrated conceptual information to be > 0, no partition should be able to divide the system into non-interacting parts (bidirectionally). Finally, according to the exclusion postulate, out of many possible constellations of concepts generated by overlapping sets of elements only one exists: the one that is maximally irreducible. Thus, one needs to evaluate Φ^{MIP} for all sets of elements s, i.e. s = A, B, C, AB, AC, BC, ABC (injecting noise in links between the set and its environment). The set of elements generating the constellation with the maximum value of Φ^{MIP} (max Φ^{MIP}, or *maximally integrated conceptual information*) constitutes the main complex within the overall system; the corresponding concept space (simplex) is called *qualia space*, and the constellation of concepts it generates – the *maximally integrated conceptual*

(information) structure - is called a *quale Q*.[11,12] An exhaustive analysis of the system in Fig. 11.4 shows that the full set ABC constitutes a complex, as no other set of elements yields integrated conceptual information structures having a higher value of Φ^{MIP}. In larger systems, one would first identify the main complex and then, recursively, identify other complexes among the remaining elements. Therefore, a *complex* can be defined as *a set of elements generating a maximally irreducible conceptual information structure*. In essence, then, just like a concept specifies a particular, maximally integrated distribution of system states out of possible distributions (a point in concept space), a complex specifies a particular, maximally integrated conceptual information structure (constellation of points) out of possible conceptual information structures in concept space. As indicated by the information axiom, that constellation differs in its particular way from other possible constellations.

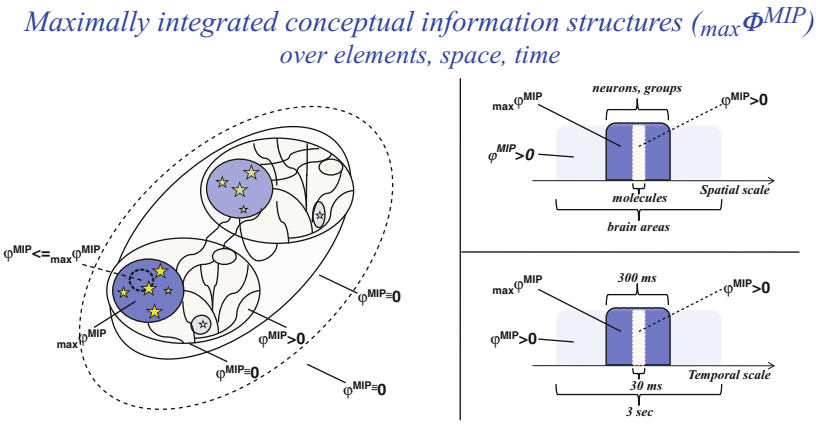
A schematic representation of a reduction of a system into complexes plus the residual interactions among them is illustrated in Fig. 11.5a. Note, for example, that due to the exclusion postulate, although complexes can interact, they cannot overlap. Thus, when two complexes of high $_{max}\Phi^{MIP}$ interact weakly, their union does not constitute a third complex, even though its Φ^{MIP} value may be >0: once again, there is no need to postulate additional entities, because they would make no further difference beyond what is accounted by the two complexes of high $_{max}\Phi^{MIP}$ plus their weak interactions.[13] This is a direct application of Occam's razor: "entities should not be multiplied beyond necessity."[14] We recognize this principle intuitively when we talk to each other: most people would assume that there are just two consciousness (complexes of $_{max}\Phi^{MIP}$) that interact a little, and not also a third consciousness (complex of lower Φ^{MIP}) that includes both speakers. In summary, a complex is an individual, informationally integrated *entity* that is maximally irreducible: (1) it cannot be partitioned into more integrated parts; (2) it is not part of a more integrated

[11] Within conceptual information, one can distinguish a backward portion (specified by the cause repertoires), or understanding; and a forward portion (specified by the effect repertoires), or control.

[12] Note that constellations of concepts must satisfy several requirements: (1) they must be physically realizable; (2) they must be self-consistent (that is, concepts that exclude/contradict each other cannot coexist; i.e., their product should never yield a distribution with zeros everywhere); (3) they must be irreducible. If these requirements are satisfied, ideally a constellation of concepts should also: (1) have as many concepts as possible; (2) they should be as irreducible as possible; (3) they should be as informative as possible about concept space, i.e., sample it as uniformly as possible (acting as representative "prototypes" of possible contingencies).

[13] Unless, of course, the interactions become so strong that $_{max}\Phi^{MIP}$ for the union exceeds that of each part, in which case the parts merge into a single complex.

[14] Occam's razor conventional formulation, "entia non sunt multiplicanda praeter necessitatem," is probably due not to Occam or his teacher Duns Scotus, but to John Ponce. It has important applications in the context of Solomonoff theory of inductive inference and compressibility (Solomonoff 1964), see also (Hutter 2005). If one can compress a wiring diagram into a product of smaller diagrams (e.g., by finding *k*-connected subgraphs) plus some residual terms, one identifies separate integrated conceptual information entities that cannot be reduced further (complexes), and beyond which no additional "higher" entities exist. Each complex is then characterized by a particular integrated conceptual information structure, within which different repertoires specified by subsets of elements exist only to the extent that they are not reducible.



- Integrated conceptual information Φ^{MIP} is the difference between the constellation (conceptual information structure) generated by the whole and that generated by its minimum information partition
- Maximally integrated conceptual information $_{max}\Phi^{MIP}$ is a local maximum of $_{max}\Phi^{MIP}$ over all sets of elements/space/time
- The corresponding maximally integrated conceptual information structure is a constellation or "shape" Q in qualia space

Fig. 11.5 Complexes: maxima of integrated conceptual information over elements, space, and time. See text for explanation

system; (3) it is separated through a boundary from everything external to it (it *excludes* it). In this view, any system of elements "condenses" into distinct, nonoverlapping complexes that constitute local maxima of integrated conceptual information.

11.7 Optimal Spatio-Temporal Grain

The exclusion postulate should be applied not only over sets of elements, but over different spatial and temporal scales. For any given system, one can group and average the states of several micro-elements into states of a smaller number of macro-elements. Similarly, one can group and average states over several micro-intervals into longer macro-intervals. For each spatio-temporal grain, one calculates CER, concepts (maximally irreducible CER), and complexes (sets of elements generating maximally integrated conceptual information structures). By the exclusion postulate, a particular set of elements, over a particular spatio-temporal grain, will yield the max value of Φ^{MIP}, thereby excluding any overlapping subsets and spatio-temporal grains.

As an example, consider the brain: over which elements should one consider perturbations and the repertoire of possible states? A natural choice would be neurons, but other choices, such as neuronal groups at a coarser scale, or synapses at a finer scale, might also be considered (not to mention molecules and atoms). Importantly, in certain circumstances, a coarser spatial scale ("macro"-level) may produce a complex with higher values of Φ^{MIP} than a finer scale ("micro"-level), despite the smaller number of macro- compared to micro-elements. In principle,

then, it should be possible to establish if in the brain consciousness is generated by neurons or groups of neurons. In this case the exclusion postulate would also mandate that the spatial scale at which Φ^{MIP} is maximal, be it neurons or neuronal groups, excludes finer or coarser groupings: informationally there is no superposition of (conscious) entities at different spatio-temporal scales (Fig. 11.5b).

Similar considerations apply to time. Integrated information can be measured at many temporal scales. Neurons can choose to spike or not at a scale of just a few milliseconds. However, consciousness appears to flow at a longer time scale, from tens of milliseconds to 2–3 s, usually reaching maximum vividness and distinctness at a few hundred milliseconds (Fig. 11.5c). IIT predicts that, despite the larger number of neural "micro"-states (spikes/no spikes, every few milliseconds), Φ^{MIP} will be higher at the level of neural "macro"-states (bursts of spikes/no bursts, averaged over hundreds of milliseconds). This is likely the case because a set of neurons widely distributed over the cerebral cortex can interact cooperatively only if there is enough time to set up transiently stable firing patterns by allowing spikes to percolate forward and backward. Again, the exclusion postulate would mandate that, whatever the temporal scale that maximizes Φ^{MIP}, be it spikes or bursts, there is no superposition of (conscious) entities evolving at different temporal scales.

Importantly, for a macro-level to beat a micro-level, despite the much larger number of states that are available to the micro-level, some features are especially important: (1) the presence of some degree of indeterminacy at the micro-level (due to intrinsic noise or to perturbations from the environment); (2) many-to-one mapping, such that many input states can produce the same output state, giving rise to irreversibility; (3) macro-mechanisms structured in such a way that they group noisy micro-states together in an advantageous manner; (4) the fact that, from the intrinsic perspective of the macro-system, all possible perturbations (i.e., counterfactuals) must be conceived as applied to macro-states. This means that the actual distribution of micro-states underlying the macro-level distribution will be different from their micro-level maximum entropy distribution, thus accounting for emergence without violating supervenience. In summary, the level at which "things" really exist in and of themselves, i.e., from the intrinsic perspective, in both space and time, is the level at which Φ^{MIP} is maximized—that is, the level at which "causal power" is maximal. In other words, what really exists (and excludes any other level) is what makes the most difference.

11.8 Identity Between Integrated Conceptual Information Structures (Qualia) and Experiences

In summary, a particular set of elements at a particular spatio-temporal scale yielding a maximum of integrated conceptual information ($_{max}\Phi^{MIP}$) constitutes a complex, a "locus" of consciousness. By definition, a complex is maximally integrated informationally, i.e., it is maximally *irreducible* to smaller conceptual information entities. The set of its concepts—maximally irreducible CERs ($_{max}\varphi^{MIP} > 0$) specified by various subsets of elements within the complex—constitute a

maximally integrated conceptual information structure or *quale* (Fig. 11.4)—a shape or constellation of points in qualia (concept) space.

Having defined complexes and qualia based on ontological postulates, IIT posits *identities* between phenomenological and informational/causal aspects of systems. The central identity is the following: *an experience is a maximally integrated conceptual information structure* or *quale*—that is, a maximally irreducible constellation of points in qualia space. According to IIT, *the maximally integrated conceptual information structure generated by a complex in a state completely and univocally determines the quality of experience*. Tentative corollaries of this identity include the following: (1) the particular "content" or quality of the experience is the shape of the maximally integrated conceptual information structure in qualia space; (2) a phenomenological distinction is a maximally irreducible cause-effect distinction (a concept). In other words, unless there is a mechanism that can generate a maximally irreducible CER (concept)—a distinct point in the quale—there is no corresponding distinction in the experience the subject is having; (3) the intensity of each concept is its $_{\max}\varphi$ value; (4) the "richness" of an experience is the number of dimensions of the shape; (5) the scope of the experience is the portion of qualia space spanned by its concepts; (6) the level of consciousness is $_{\max}\Phi^{\mathrm{MIP}}$—the maximally integrated conceptual information; (7) the similarity between concepts is their distance in Q, given the appropriate metric; (8) clusters of nearby concepts form modalities and submodalities of experience; (9) the similarity between experiences would be given by the similarity between the corresponding shapes (see also the final section and Tononi 2008, 2013), and so on.

In principle, then, given the "wiring diagram" and present state of a given system, IIT offers a way of specifying the maximally integrated conceptual information structure it generates (if any).[15] According to IIT, that structure completely

[15] The complete characterization of an experience or quale would thus require specifying all of the concepts (cause-effect repertoires in Q) of a complex. From the intrinsic perspective, these concepts provide the information necessary to distinguish that experience from any other. From the extrinsic perspective, knowing these distributions and their degree of irreducibility, one would know all there is to be known about that experience. It is interesting to ask how much information that is (in terms of algorithmic complexity or incompressible information). Clearly, the input–output matrix of a system (or transition probability matrix TPM), if known and available to perform manipulations (injecting noise), could be used to derive all the quantities discussed here. However, the information in the TPM is both uncompressed and implicit. It can be an *uncompressed* TPM when a large TPM reduces to the product of the smaller TPMs, as indicated by $\varphi^{\mathrm{MIP}} = 0$. More generally, finding $_{\max}\varphi^{\mathrm{MIP}}$ and $_{\max}\Phi^{\mathrm{MIP}}$ over subsets of elements would indicate how best to compress a large TPM into the product of smaller, maximally irreducible TPMs, plus some extra terms. Also, it may turn out that a TPM at the finest spatio-temporal grain may be compressed to a coarser spatio-temporal grain with no loss (or indeed gain) in information. This aspect is captured again by finding $_{\max}\Phi^{\mathrm{MIP}}$ over different spatio-temporal scales. The TPM is also *implicit*: while it contains all the information necessary to find complexes and specify their quale, making them explicit requires work. One must extract the repertoires specified by each element and subset of elements, find the MIP to establish which subsets integrate information, which sets of elements are maximally irreducible (concepts and then complexes), and at which spatio-temporal grain size. This requires examining the effects of a large number of perturbations (performing partitions and injections of noise/max entropy) within a large combinatorial space. At a minimum, one would need to calculate probability distributions specified by each element, from which one can calculate all the distributions specified by subsets of elements (as the product of distribution at lower levels in the power-set). From this one can establish, through

specifies "what it is like to be" that particular mechanism in that particular state, whether that is a set of three interconnected logical gates in an OFF state; a complex of neurons within the brain of a bat spotting a fly through its sonar; or a complex of neurons within the brain of a human wondering about free will. In the latter examples, the full integrated conceptual information structure is going to be extraordinarily complex and practically out of reach: we are not remotely close to having the full "wiring diagram" of the relevant portions of a rodent or human brain; even if we did, obtaining the precise quale would be computationally unfeasible.[16] Nevertheless, by comparing some overall features of the shapes of qualia generated by different systems or by the same system in different states, it would be possible to evaluate broad similarities and differences between experiences. IIT also implies that, if a collection of mechanisms does not give rise to a single maximally integrated conceptual information structure, but to separate qualia each reaching a maximum of integrated conceptual information, then there is nothing it is like to be that collection, whether it is an array of electronic circuits, a heap of sand, a swarm of bats, or a crowd of humans.

11.9 Matching

So far, the maximally integrated conceptual information structures generated by a system of elements have been considered in isolation from the environment—as is the case for the brain when it dreams. But of course it is just as important to consider how integrated conceptual information structures are affected by the external world, especially since the mechanisms generating them become what they are through a long evolutionary history, developmental changes, and plastic changes due to interactions with the environment.

In any situation, a complex of high ${}_{\max}\Phi^{\mathrm{MIP}}$ has at its disposal a large number of concepts—maximally irreducible CERs specified within a single conceptual information structure. These concepts allow the complex to understand the situation and act in it in a context-dependent, valuable fashion (and thus to answer a large number of questions, see below). *Cause-effect matching* (*M*) measures how well the integrated conceptual information structure generated by an adapted complex fits or "matches" the cause-effect structure of its environment. It can be expressed as the average change in the average conceptual information of the quale generated by

appropriate partitions, which subsets specify maximally irreducible points and, finally, which maximally irreducible subsets constitute complexes. It would be interesting to know if the most economical characterization (e.g., algorithmic complexity) of a particular conceptual information structure would correspond to the minimal set of causal processes generating it. In this case, obtaining complexes and their quale (integrated conceptual information structure) would be equivalent to finding the most compressed description of the causal structure of a physical process.

[16] In any case, *describing* a quale would not be the same as *being* that quale.

a complex C when it interacts with its environment, compared to when it is exposed to uncorrelated noise (or structureless environment):

$$\langle M \rangle = \left\langle {}_{\max}\Phi^{\mathrm{MIP}}(C_{\mathrm{World}}) \right\rangle - \left\langle {}_{\max}\Phi^{\mathrm{MIP}}(C_{\mathrm{Noise}}) \right\rangle$$

In the course of evolution, development, and learning, one would expect that the mechanisms of a system change in such a way as to increase matching. For example, since cause-effect information/integrated information are the minimum between the input and output sides, these would have to be "well-matched." That is, a state s that specifies a cause repertoire with high ${}_{\max}\varphi^{\mathrm{MIP}}$ $(P \mid s)$ should also specify a paired effect repertoire with high ${}_{\max}\varphi^{\mathrm{MIP}}$ $(F \mid s)$. This would ensure a high flow of irreducible cause-effects between the input and output sides when averaging over the distribution of states occurring during interactions with the environment. Moreover, the interactions with the environment would have to match not only the input and output values of ${}_{\max}\varphi^{\mathrm{MIP}}$, but also to match specific cause repertoires with specific effect repertoires in a way that yields perception-action cycles of high adaptive value: in short, the "right" cause should lead to the "right" effect. Moreover, since $<M>$ depends on ${}_{\max}\Phi^{\mathrm{MIP}}$, optimization of $<M>$ means that concepts should not only have high average ${}_{\max}\varphi^{\mathrm{MIP}}$, but there should be *many different concepts*, which yield qualia with high conceptual information, thereby avoiding redundancy. If well matched, large qualia provide a broad context to understand a situation and to plan an appropriate action.[17]

In general, then, one can expect high matching if the complex can deploy a large number of different, highly specific concepts, of many different orders (subsets of 1, 2, … n elements), that jointly capture many aspects of the causal structure of that environment. In this way, information about the environment is efficiently distributed to many subsets of a complex, each of which is specialized for different features, and these jointly lead to highly specific, valuable effects. On the other hand, one expects low matching if information from the sensory input and to the motor output is not distributed efficiently—say because a system is not integrated but organized into parallel channels. Matching will be low also if the elements of a system are not specialized—say due to a homogeneous connectivity that would force all elements to perform the same operation and generate little integrated information.[18] In conclusion, based on theoretical considerations and supported

[17] Within matching, one can distinguish a backward portion (specified by the cause repertoires), or *representation capacity*; and a forward portion (specified by the effect repertoires), or *action capacity*.

[18] Note that in an unpredictable environment it is important not only to have a large repertoire of possible actions, but also to have many different ways of achieving the same effect, i.e., degeneracy (Tononi et al. 1999). High degeneracy implies both high effective information and high integration in the forward repertoire component of the concepts available to a complex. In general, if information integration is high, a small subset of elements within a complex should be able to affect many other elements (pleiotropy). At the same time, many subsets of elements should be able to produce the same effect over a small subset of outputs (degeneracy).

by simple simulations (Tononi, Sporns and Edelman 1996, 1999), it is expected that matching should increase when a system adapts to an environment having a rich, integrated causal structure. Moreover, an increase in matching will tend to be associated with an increase in information integration and thus with an increase in consciousness.[19]

11.10 Information and Causation

The framework presented above will certainly need to be expanded and refined. However, even in its current from, it can shed some light on some broad theoretical issues that assume critical relevance if one takes integrated information to be a fundamental, intrinsic feature of reality (Tononi 2008).[20] One of these concerns the relationship between information and causation.

IIT assumes that mechanisms in a given state are *intrinsically* associated with certain maximally integrated conceptual information structures, which they specify irrespective of external observers. Each integrated information structure is specified if and only if the "cause-effect" mechanisms are in working order and can "choose" among alternatives, that is, select a particular subset of past and future states that are compatible with their present state. Moreover, a concept (point) in an integrated conceptual information structure exists if and only if it is maximally irreducible to subconcepts. Finally, an integrated conceptual information structure only exists if it constitutes a maximum of integrated conceptual information over elements, space, and time.

From these premises, it is worth considering more closely the relationship between information and causation. Causation has often been interpreted as a correlation between successively observed events, as pointed out by David Hume: by observing that event 1 is reliably followed by event 3, we infer that 1 causes 3. This view of causation as strength (*reliability*) of a correlation is akin to the traditional view of information from the extrinsic perspective, as in Shannon's formulation, where a correlation between 1 and 3 means that one event carries information about the other (mutual information). Some more recent formulations, such as transfer entropy, impose the additional criterion of the directionality of prediction. However, it would seem that, to assess causation, it is not enough to observe a system, but it is necessary to perturb it and see what happens. In this vein, Judea Pearl has developed an interventional or perturbation-based model of causation: for instance, one does not merely observe the sequence 1, 3, but one imposes input state 1 and sees whether

[19] This is because $<M>$ is bounded by $<_{max}\Phi^{MIP}>$.

[20] Since consciousness undoubtedly exists (indeed, it is the only thing whose existence is beyond doubt), if each individual consciousness is an integrated conceptual information structure, then integrated information must be a fundamental ingredient of reality—as fundamental as mass, charge, or energy (Tononi 2008).

event 3 is reliably observed (while the opposite may not be true). In this case one can conclude that 1 caused 3, going beyond a mere correlation.

Conceptualizing causation properly also requires the consideration of *counterfactuals*, that is, what would have happened if instead of event 1 some other event had occurred. For instance, would effect 3 still have happened if, instead of imposing 1, one had imposed perturbation 4, 5, 6, and so on? If it turned out that the system always ends up in state 3, we would begin to think that 3 was not so much caused by the preceding state 1, but rather, that 3 was inevitable. In other words, it would seem that, the less a cause is specific, the less of a cause it is. Some further thought indicates that properly considering counterfactuals ties the notion of causation even more closely with that of information, precisely because it implies *specificity*. In the general case, it would seem that one should consider all possible counterfactuals. That is, one should perturb the system with all possible initial states (the maximum entropy distribution) and see what it does. This is exactly what is done by measuring cause-effect information as defined above. CEI certainly depends on the *reliability* of the effects of a particular perturbation, as it decreases with noise. Cause-effect information also depends critically on *specificity*: it is high if only some out of many past perturbations can give rise to the present effect.

If cause-effect information can indeed capture some aspects of causation—namely reliability and specificity—what is the relation between causation and integrated information—the extent to which effective information is irreducible, as established from the intrinsic perspective of a system? As was argued above, if a candidate CERs, as measured by CEI > 0, can be reduced to the product of independent components, as indicated by $\varphi^{\text{MIP}} = 0$, then there is no reason to posit its existence as an additional mechanism, because there are no further cause-effects to be accounted for beyond those accounted for by component mechanisms. In other words, true causation requires not only that CEI > 0, but also that $\varphi^{\text{MIP}} > 0$.

An even stricter notion of cause is imposed by considering the notion of maximal integrated information. As was also argued above, once an element is in a certain present state (say ON), it makes no difference which of the possible causes of its being ON may have occurred, so one can simply consider the maximally irreducible set of past causes and future effects—those that make most of a difference (MICE). This notion of maximally irreducible causation, enforced by the exclusion postulate, has the virtue of avoiding the paradoxes posed by multiple causation (causes should not be multiplied beyond necessity).

Based on the same notions one further identifies sets of elements—complexes—that specify a maximally irreducible integrated conceptual information structure ($_{\text{max}}\Phi^{\text{MIP}}$) at an optimal spatio-temporal scale. The equivalence between maximal information integration and maximally irreducible causation means that causes intrinsic to a complex, at the optimal spatio-temporal scale, *supersede* external causes or causes acting at lower or higher spatio-temporal scales. From the extrinsic perspective, one can certainly describe an element belonging to a complex as subject to influences both intrinsic and extrinsic to the complex. However, from the intrinsic perspective, the causes inside the complex constitute the most irreducible set of causes—the one that most accounts for the behavior of that element, and

there is no need to consider extrinsic causes.[21] Similarly, from the extrinsic perspective one can certainly describe the behavior of elements within a complex both in terms of micro-causes (at the level of micro-elements) and of macro-causes (at the level of macro-elements): both can be useful perspectives on how the system functions (although one will be the most informative). However, things are different from the intrinsic perspective—that of the complex itself. If a set of macro-causes accounts for the behavior of a complex better than a set of micro-causes, i.e., if the micro-level is causally less complete than the macro-level, then the exclusion postulate (Occam's razor) implies that the behavior of the system is determined by the macro-causes only, and the micro-causes should not be double-counted.[22]

Finally, it follows that, if consciousness itself can be identified with a maximum of integrated information, then *consciousness is supremely causal*—an integrated structure of causation that supersedes any lesser causes.[23]

11.11 Free Will and Irreducibility

These considerations have direct bearing on the issue of free will. A first consequence of the previous account of causation as integrated information is that freedom of will is above all an issue of irreducibility of choice, rather than of indeterminacy.

Traditionally, one recognizes that freedom requires autonomy from extrinsic constraints—if a choice is forced upon us by external causes—typically causes in the environment and not in our own head, then our freedom is reduced. More recently, it has become apparent that even factors within our own head may reduce

[21] It is interesting to consider how the notion of maximally irreducible set of past causes of future effects maps onto accounts of trajectories of dynamical systems, for example accounts of how an element may be enslaved by one of two weakly coupled attractors, though being subjected to causal influences from both. More generally, it is interesting to consider how the intrinsic notion of causation indicated here maps onto an extrinsic notion of causation developed along parallel lines (Hoel et al., in preparation). In the extrinsic perspective, one takes a given event (i.e. an observed state) and considers what past event actually caused it (as opposed to what could have potentially caused it, as in the intrinsic view) and what are its actual future effects (as opposed to potential effects). In this way, it is possible to define an extrinsic notion of cause-effect power based on the sufficiency (reliability) and necessity (specificity) of the mechanisms mediating the transition from one event to the next, and the size of the repertoire of counterfactuals. By applying exclusion, one can then proceed to partitions to identify maximally irreducible ("core") cause-effects as well as sets of cause-effects ("cause-effect complexes").

[22] That is, one should not double-count intrinsic causes, just as one should not double-count information. In terms of dynamical systems, this means that micro-variables are "enslaved" by macro-variables.

[23] In this sense, integrated information can be said to be a measure of intrinsic causation. And a complex—defined from the intrinsic perspective as a maximally irreducible set of maximally irreducible cause-effect repertoires (concepts)—can be said to be truly *causa sui*.

our freedom—for instance, an action that is triggered by neural automatisms, as in certain psychomotor seizures, or a pathological compulsion, as in Tourette's syndrome. While originating in our own brain, such actions are perceived as alien or "ego-dystonic," and thus as reducing our freedom just as much as if they were forced from the outside. In short, it is clear that freedom of will requires *autonomy* from constraints extrinsic to our conscious self, whether these constraint are in the environment or in "alien" parts of our own head, parts that are beyond our control.

What is often not realized, however, is that autonomy is only one aspect of what we cherish about freedom. A reflex arc that can merely choose whether to blink or not, or any simple, isolated mechanism going through a limit cycle of just a few states, may be perfectly autonomous, but it does not seem to be particularly free. On the other hand, a human being agonizing over a moral dilemma, who tries to deliberate according to his conscience, his beliefs and values, his knowledge and understanding of how things are, his character, history, memories, aspirations, and feelings, as well as those of others within the circle of his empathy, in short, who tries to choose according to a large set of concepts that live within his present conscious experience—the tribunal of consciousness—is the referent of choice for freedom, the best example we have of what we mean by the exercise of free will.

Based on the framework presented above, such a situation obtains when a choice is made in the context of a large, maximally irreducible integrated conceptual information structure—the quale generated by a complex having high $_{\max}\Phi^{\mathrm{MIP}}$. It follows that, from the intrinsic perspective and given the equivalence between integrated information and causation, the more a choice is determined from inside a complex, the more it is free. Thus, if how the present state of a system determines its future states based on its past ones—its choice—can be reduced to a simple mechanism—a simple causal concept constituting a minimal quale (if light, then blink), then that choice is worth only a few bits of integrated information/causation. If instead a choice can be accounted only by considering the joint consequence of a large number of irreducible causal concepts, within a maximally irreducible quale, then that choice is worth very many bits. In this view, then, *maximizing freedom is not just minimizing external constraints, but maximizing internal cause-effects*—cause-effects expressed by a maximally irreducible conceptual information structure. Other, equivalent ways to express this conclusion are as follows. First, if a conscious experience is indeed a maximally integrated conceptual information structure, and the more one is conscious, the larger is the number of distinct concepts in that structure, then a choice is the freer the more it is conscious: *more consciousness, more freedom*. Second, and a bit paradoxically, *a choice is the freer, the more it is determined* (intrinsically). That is, it is the freer, the less it can be reduced to a set of simple cause-effects, but it requires a large integrated conceptual information structure to account for it. This is one fundamental sense in which the key notion of *alternative possibilities*—the feeling that one could have acted otherwise—which is essential to the feeling if being responsible for one's action, is captured by a large maximally integrated conceptual information structure: such a structure implies a very large number of counterfactuals (alternative possibilities) that are under the control of the agent (they are part of his consciousness).

11.12 A Neurophysiological Example: Reflex and Conscious Actions

The above points about information and causation in a complex can also be helpful when considering neurophysiological experiments, especially given the interest concerning electrophysiological and neuroimaging approaches addressing decision making and its relationship to free will.

A simple but useful contrast is to compare a response mediated by a conscious corticothalamic main complex with one mediated by a reflex arc. Say the task is to blink if a light turns on and not to blink if it turns off. For a reflex arc—say one producing a blink in response to the light—the underlying wiring diagram includes just a small chain of neurons and connections. The corresponding maximally integrated conceptual information structure would be equally small—indeed just a simple concept, and it would carry hardly any experiential quality. For a conscious human performing the same task, instead, the relevant wiring diagram would be vast, including a large portion of the corticothalamic system. The corresponding maximally integrated conceptual information structure would be extraordinarily large and complex, containing a huge number of distinct points (concepts). This quale would correspond to the experience of seeing the light, and may also include, in a context-dependent manner, the intention to blink, or try and suppress the blink, or to interrupt the experiment, and so on.

This complexity may be ignored when examining how the task is performed from an extrinsic perspective, say that of a neurophysiologist looking for the neurons that are activated when performing the task: one may single out a causal chain "inscribed" on top of the corticothalamic complex and represented by the neurons that fire, from a photoreceptor in the fovea to a motoneuron driving the blink, while ignoring the rest of the system. However, what is missed in such an extrinsic, observational approach, is the large set of counterfactuals. In the case of the corticothalamic main complex, as opposed to the reflex arc, the silent neurons matter: if they had fired, in any of innumerable combinations, rather than having remained silent, the output would have been different. In other words, in a complex, it is just as important that some neurons fire as that the others do not, whereas in the reflex arc there are no other neurons that could affect the end result. The tendency to consider that only neurons that fire "cause" effects, or generate information, is natural enough, but it is insufficient when dealing with an integrated system. By applying perturbations to the corticothalamic system, it would become apparent that the "causal funnel" (i.e., receptive field or cone of influence) of a neuron of the main complex ultimately leading to a voluntary blink involves the entire main complex: in other words, its output might have been different not only if the neurons prior to it in the cause-effect chain that had fired had instead not fired, but also if neurons that were silent had instead fired.

Similar considerations can be applied to classic and more recent experimental results showing that the brain comes to a decision hundreds of milliseconds before a subject becomes conscious of that decision, (Libet et al. 1991, Fried et al. 2011)

which he interprets as his own, and as having been made of his own free will. In some cases, one can find brain activity traces that predict a decision above chance several seconds before the action (Soon et al. 2008). It is also possible to fool subjects into thinking that they willed an action that was instead chosen beforehand by the experimenter (Wegner 2003). It is even conceivable to produce the "illusion" of free will through microstimulation of appropriate brain circuits. In a sense, these demonstrations are not surprising. Knowing a person beliefs and circumstances allows an external observer to predict a willed action well above chance. Moreover, it is clear that any decision must ultimately be taken by mechanisms in a state, i.e., by brain circuits. But is free will then merely an illusion, a conscious feeling of responsibility that is misguided?

Let us briefly consider three possible arguments. First, since the decision is actually due to brain circuits, it is the brain circuits that are responsible—not myself of my own free will. As with the example of the reflex blink compared to the conscious decision to blink, it all depends on which circuits are responsible. If the circuits involved form a mere reflex arc, then the "decision" is unconscious and it is not truly "my" decision. If the decision cannot be ascribed to anything less than the main complex giving rise to my experience, then the decision is indeed mine and maximally irreducible. As we have seen, a very large number of concepts are involved even in the mere decision "to blink or not to blink," many of which could change the decision depending on the context (that is, they instantiate counterfactuals). Saying that the decision is due merely to the neurons the neurophysiologist is paying attention to, those that increase their firing just before my decision, is to take the tip for the iceberg.

Second, an external observer, such as a neurophysiologist, could under certain circumstances predict the decision with above chance accuracy in advance of the subject realizing his decision, simply by monitoring relevant neural circuits. Again, this simply means that, by localizing the necessary final stages where the decision to act or not is ultimately implemented—the tip of the iceberg—an external observer can often anticipate the decision with some accuracy and before a fully developed "attractor" involving the entire main complex is established. However, by the same token, the external observer may have no idea whether the same decision has been taken mostly "locally" or depending on a much wider context—for example, I may have decided that I should interrupt the experiments having realized that I should rush home. Moreover, as is the case for myself, the inherent indeterminacy of the decision will in any case preclude perfect accuracy also for an external observer (see below).

Third, a subject can be fooled into thinking that an externally determined decision is his own. In this respect, it is worth pointing out that, based on the account above, the core concept specified by a particular circuit—say a group of neurons that, whenever activated by its sources, produces a certain kind of outputs—is constituted by the maximally irreducible set of causes of its firing and its effects (MICE). The core concept is what is postulated to contribute to consciousness, yielding the conscious concept "I willed it." The implication is that since other influences that might have caused that group of neuron to fire—for

example extrinsic microstimulation—do not belong to its MICE, consciousness interprets the activation based on its core concept—the most parsimonious interpretation—which is that the decision was willed by the self. The source of the decision in this case is an illusion—but a perfectly justifiable one, just as is the case with many visual illusions. But the existence of illusions of will, just as that of visual illusions, does not imply that free will in general is illusory, or that visual experience is unreal.

11.13 Free Will and Indeterminism

This view implies that a choice is free to the extent that it cannot be reduced to the action of external constraints, and to the extent that the internal constraints that determine it cannot themselves be reduced. In other words, a choice is free if it cannot be accounted for by anything less than one's whole conscious, deliberating self, and it is the freer the larger that conscious self is that is brought to bear on that choice (in terms of the irreducibility of the maximally integrated conceptual information structure that constitutes one's conscious experience relevant to that choice). This notion captures an essential aspect of what we intend by free will, and perhaps all we need, at least according to compatibilists—those who think that free will can be compatible with determinism. And yet many people would still feel that, without some indeterminism, their choice would be pre-ordained and therefore not truly free—it would negate the essential feeling that "I can decide this way, but I could also have decided otherwise." In other words, while the huge number of counterfactuals implied by a large maximally integrated conceptual information structure does in fact offer a number of alternative possibilities, and the one that is eventually chosen cannot be reduced to anything less than the joint action of all the constraints (concepts) implemented within one's consciousness, including one's beliefs, feelings, goals, character, and history, nevertheless it feels that, if the choice could not have come out any different, then the freedom we associate with responsibility is an illusion. For when one deliberates, one is merely, as it were, running the computation that will determine the outcome. It is true that the computation can be run by nothing less than oneself, and that its outcome cannot be reached cheaply without going through all the complicated interactions that implement it. But if the results could not have been turned out differently, then we are just computing the answer, not choosing it. That is, it may have seemed as if there were innumerable alternative possibilities before the computation, but given a deterministic computation, there was really no alternative to the choice that was actually made.

For a long time, then, incompatibilists have searched for ways in which some degree of indeterminism might come to the rescue and salvage the notion that free will is not merely an illusion. Indeterminism seems intrinsically warranted at least at the micro, quantum level of reality, and may also be warranted in practice due to the impossibility to predict perfectly the evolution of large, complex systems without actually "running" them. For just as long, however, critics such as David Hume

have pointed out that merely adding some measure of randomness to the processes leading to a choice cannot possibly help. This is because a sprinkling of randomness on a giant clockwork mechanism within one's brain in no way makes one more responsible of a choice; in fact, it makes one less responsible. In other words, the kind of freedom offered by some random process is not a freedom of will worth wanting—it merely adds an element of arbitrariness.

This conclusion seems inescapable, but is it warranted? We will not worry whether indeterminism is due to fluctuations that are intrinsic to the mechanisms underlying one's consciousness, or due to unpredictable input from extrinsic sources—this makes no difference to the system itself, as long as it is unpredictable. Instead, we will briefly examine the role of indeterminism at the micro-level in making possible the emergence of macro-levels of information integration, both in time and in space. As was mentioned above, IIT claims that a maximally irreducible conceptual information structure obtains among a particular set of elements (a complex), defined over the temporal and spatial grain at which Φ^{MIP} reaches a maximum. For our own consciousness, for example, it is likely that Φ^{MIP} reaches a maximum over a few hundreds of milliseconds, which is the time necessary to establish irreducible interactions among widely distributed elements in the cerebral cortex. It is also likely that the elements over which Φ^{MIP} reaches a maximum are macro-elements, such as neurons or even groups of neurons, rather than individual atoms or subatomic particles. As we also saw, however, the emergence of macro-time scales requires among other things some randomness at the micro-level, otherwise the macro-level cannot beat the micro-level in terms of how much integrated information it generates. The implication is that, for the macro-level in space and time upon which our consciousness depends to have more causal power than the underlying micro-level—for the system to be progressively more causal, i.e., determined—it is necessary for it to be fundamentally unpredictable, due to unpredictability at the micro-level. That is, while a choice is the freer the more it is (intrinsically) determined, it can never be completely determined, preordained, or predictable, not from the outside and not from the perspective of the agent himself. One way of seeing this is to think that, as much as one can try to make a decision by marshalling as many factors (concepts) as possible within the tribunal of consciousness; as much as one may try and influence the scaffold of the maximally integrated conceptual information structure that will make future decisions; and as much as one may try to be consistent in one's decisions under similar circumstances; nevertheless, if consciousness emerges at a macro-level in time and space, which definitely seems to be the case, then it is not at all fully determined that the decision will be the same under the same "macro"-circumstances.

An indication of how this view differs from the usual way to frame the dilemma between determinisms as eliminating freedom, and indeterminism as merely substituting chance for responsibility, can be obtained as follows: some indeterminism is a given, and it is in fact an essential ingredient for macro-levels of organization to supersede lower levels, as seems to be the case for our consciousness; on the other hand, growing from this inescapable background of micro-level indeterminism, consciousness emerges as a complex edifice of maximally integrated

constraints (concepts) that increase the determination of our choices as much as possible, but never completely. That is, freedom of will is a fight in which order (integrated information) minimizes disorder (lack of constraints) by taking into account as many constraints (knowledge) as possible. A bit like building a society or a civilization out of relative chaos, or a bit like evolution creating macro-order out of micro-level disorder, thus increasing complexity. But as with societies, civilizations, and evolution, what will actually occur can never be predicted exactly before it happens, and micro-fluctuations—a queen and a squire falling in love, two lizards separated from the mainland after a flood—may initiate an extraordinary turn of events that nobody could predict, not even the universe itself.

11.14 Conclusion

In summary, IIT provides a useful framework for addressing some of the classic issues surrounding the problem of free will.

The requirement for *autonomy* implies that, to be free, one must be independent from constraints outside one's deliberating consciousness. These include both environmental constraints, such as limitations that force us to a particular choice or that impede our own choice, and unconscious, "alien" constraints that, while generated somewhere within our brain, affect our actions largely outside the control of the conscious self. IIT guarantees this requirement by considering the choices available to a complex of elements from its own intrinsic perspective—independent of external constraints. Since a complex is a maximally irreducible entity that constitutes a private, individual consciousness, a choice can be free only to the extent that it is decided within the tribunal of one's own consciousness.

The requirement for *understanding* implies that, to be free, a choice must be based on a concept of what is at stake—I am freely choose between right and wrong only if one has a notion of which actions are right and which are wrong under some circumstances. IIT clarifies the notion of a concept—how the present state of a mechanism within an integrated conceptual information structure specifies past causes and future effects—the CER. Thus, the concept of "right"—a high-order invariant—requires an irreducible mechanism that groups certain situation-actions available to a complex of elements within the category "right" and rules out other situation-actions as being "not right." Only if I have such a concept can I be responsible for my choice.

The requirement for *self-control* implies that, to be free, one must be able to influence one's choices. That is, merely registering some state of affairs but not being able to influence its outcome does not allow for freedom. IIT prescribes that a system that could categorize its own past states without any ability to affect its own future states would not form a complex.

The requirement for *alternative possibilities* implies that a choice can be free only to the extent that alternative choices are available. Thus, a system that can only distinguish among a few past states, and choose among a few future states, cannot

be very free. On the other hand, a complex that maximizes Φ^{MIP}—a highly conscious one—is one that can choose a particular situation-action pair out of a large repertoire of alternatives. For such a complex, each choice is highly informative and thereby causal, where causation requires both reliability and specificity (cause-effect information).

The requirement for *irreducibility* implies that a choice can only be free if it cannot be ascribed to anything less than oneself—I am the only entity that can be said to be *responsible* for the choice. That is, when asking who is responsible for the choice, the answer should be "me," meaning *all* the circuits underlying my present conscious experience, and nothing less than that. For example, while I may have a general concept of right or wrong, I also have many other concepts, both general, say of what is advantageous to me, and particular, such as the specific circumstances in which the choice is made, all of which are necessary to understand the context. Crucially, the choice to act one way rather than another could be swayed by such circumstances. It follows that only a consideration of the entire set of concepts within the quale is sufficient to account for the choice. IIT indicates that each experience is a maximally integrated conceptual information structure generated by a complex, and therefore what it will choose given a particular present state cannot be ascribed to anything less than the full structure, with all its concepts. This structure is maximally causal and it is both necessary and sufficient to account for the choice—anything less won't do, and nothing more is needed. Thus, each choice is a choice of the whole complex, not reducible to a number of choices made within nearly independent modules, each in a limited context. As we have seen, causation requires not only reliability and specificity, as measured by cause-effect information, but also irreducibility, as measured by integrated information. In short, IIT prescribes that a choice is the freer, the more it is caused, viz. the more it is conscious.

The requirement for *indeterminism* implies that, even though a choice may feel free by satisfying all the above conditions, if we knew for certain that a choice is completely preordained due to absolute determinism, we would conclude that the feeling of responsibility is an illusion. The choice would indeed be autonomous—ours and nobody else's; would indeed require broad understanding; would indeed be associated with the capacity for self-control; would indeed select one out of many possibilities; would indeed be irreducible to any lesser mechanisms that the one generating our own consciousness; but though fully and irreducibly, consciously ours, it would also be inescapable. Many concede that some degree of indeterminism is essentially guaranteed, not only due to quantum phenomena but simply to the unpredictability of the environment. Therefore, especially when the odds are close for some alternative possibilities, a choice may not be preordained.

However, at least since David Hume, it has been argued that this kind of indeterminism does nothing to assuage the feeling that responsibility is ultimately illusory: to the extent that a choice is determined, ultimate responsibility remains an illusion, and to the extent that it is indeterminate or random, it becomes merely

arbitrary. In this regard, IIT offers a different perspective. We know that our own consciousness does not flow over micro-time steps over micro-elements such as subatomic particles, but over hundreds of milliseconds over neurons or maybe neuronal groups. IIT claims that this is so because the underlying integrated conceptual information structure reaches a maximum at that particular macroscale in time and space. Importantly, it also claims that more integrated information can be generated at the macro-level only if there is some indeterminism at the micro-level. In this way indeterminism is never eliminated, but controlled thanks to the establishment of macro-structures that increase integrated information by "enslaving" the micro-levels. Thus, according to IIT, the correct way of considering indeterminism is not as a useless trick to try and instill a drop of freedom into a preordained cascade of mechanisms—by decreasing their causal powers. Rather, indeterminism provides a backdrop of ultimate unpredictability against which macro-level, integrated mechanisms fight to increase understanding and control—a fight for increasing the causal powers of consciousness, and the more these increase, the more freedom increases. But since this is a battle against a backdrop of indeterminism, its results are never completely predictable.

Finally, one should at least mention the requirement for *self-formation*—the idea that we become responsible for an action also by willing to change our own mechanisms, as one does when one makes a difficult, morally forming decision and tries to stick with it for life. (Kane 2005) In view of what has been said so far, it should be clear that, due to the very fact that the mechanisms underlying consciousness are themselves plastic, we can change our future self by acting in the present.

Acknowledgements Part of the material presented here is derived from the previous publications, especially Tononi, G. Integrated Information Theory of Consciousness: An Updated Account, Archives italiennes de Biologie, 2012. I thank Chiara Cirelli, Lice Ghilardi, Christof Koch, Barry van Veen, Virgil Griffith, Atif Hashmi, Erik Hoel, Matteo Mainetti, Melanie Boly, Andy Nere, Masafumi Oizumi, Umberto Olcese, and Puneet Rana for many helpful discussions and for developing the software used to compute integrated conceptual information structures (M. Oizumi, A. Nere, A. Hashmi, U. Olcese, P. Rana). This work was supported by a Paul Allen Family Foundation grant and by the McDonnell Foundation.

References

Bennett, C. H., Gacs, P., Li, M., Vitany, P. M. B., Zurek, W. H. (1998). Information distance. *IEEE Transactions on Information Theory, 44*, 1407—1423.

Fried, I., Mukamel, R., & Kreiman, G. (2011). Internally generated preactivation of single neurons in human medial frontal cortex predicts volition. *Neuron, 69*(3), 548–562.

Hutter, M. (2005). *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Berlin: Springer.

Kane, R. (2005). *A contemporary introduction to free will*. New York: Oxford University Press.

Libet, B., et al. (1991). Control of the transition from sensory detection to sensory awareness in man by the duration of a thalamic stimulus. The cerebral "time-on" factor. *Brain, 114*(Pt 4), 1731–1757.

Solomonoff, R. J. (1964). A formal theory of inductive inference. *Information and Control, 7*(2), 224–254.

Soon, C. S., et al. (2008). Unconscious determinants of free decisions in the human brain. *Nature neuroscience, 11*(5), 543–545.

Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *Biology Bulletin, 215*(3), 216–242.

Tononi, G. (2010). Information integration: Its relevance to brain function and consciousness. *Archives italiennes de biologie, 148*(3), 299–322.

Tononi, G. (2013). Integrated information theory of consciousness: An updated account. Archives italiennes de biologie, in press.

Tononi, G., Sporns, O., & Edelman, G. M. (1996). A complexity measure for selective matching of signals by the brain. *Proceedings of the National Academy of Sciences of the United States of America, 93*(8), 3422–3427.

Tononi, G., Sporns, O., & Edelman, G. M. (1999). Measures of degeneracy and redundancy in biological networks. *Proceedings of the National Academy of Sciences of the United States of America, 96*(6), 3257–3262.

Wegner, D. M. (2003). *The illusion of conscious will*. Cambridge: MIT.

Chapter 12
On Habit Learning in Neuroscience and Free Will

Javier Bernácer and José Manuel Giménez-Amaya

Abstract The notion of habit learning in Neuroscience implies the automation of an action, which thus discharges consciousness from the supervision of its performance and eventually restricts flexibility. It has also been assumed that habit learning is against free will, as it has been suggested for pathological conditions such as obsessive-compulsive disorder. This point of view, which might be controversial with other notions of habituation, could be an interesting context to analyze at what extent human actions emerge from free will and are consciously carried out. The well-known experiments performed by Benjamin Libet and replicated by others have led some scientists to deny the concept of free will in the human being. However, we think that these experiments posit further questions that should be tackled from a broader point of view. For example: does the readiness potential univocally point to the initiation of any kind of action? Can it be also found in non-deterministic novel actions? Is it causally related to the action, or is it just a "mental rehearsal" of the action to come? In this contribution, we will try to make a note on these topics in order to explain the neuroscientific concept of habit learning and to relate it to free will in a broader and more philosophical interdisciplinary framework.

Keywords Libet • Habit learning • Neuroscience • Basal ganglia • Habit • Aristotle • Thomas Aquinas

J. Bernácer (✉)
Mind-Brain Project, ICS (Institute for Culture and Society), University of Navarre, Navarre, Spain
e-mail: jbernacer@unav.es

J.M. Giménez-Amaya
Mind-Brain Project, Research Group of Science, Reason and Faith (CRYF),
ICS (Institute for Culture and Society), University of Navarre, Navarre, Spain
e-mail: jmgimenezamaya@unav.es

A. Suarez and P. Adams (eds.), *Is Science Compatible with Free Will?: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience*,
DOI 10.1007/978-1-4614-5212-6_12,

12.1 Neuroscience and Free Will

Modern scientists have realized that the big questions of life must be tackled re-visiting a classical, holistic approach. The extreme specialization in independent and reduced fields of study is evolving to interdisciplinary groups, where their highly specialized members communicate with each other to find an overall solution to the problem they are dealing with.

The first difficulty these groups may find is terminology. Some concepts are common to various disciplines, although they can have different meanings and, moreover, very distinct nuances. Such is the case of the terms "habit" and "habit learning", crucial processes for human behaviour according to philosophy and neuroscience. But, do these two disciplines agree on what a "habit" is? This chapter is intended to clarify the main features of habit learning for neuroscience, especially those referring to free will, and to show how philosophers have a different point of view when they use the term "habit". In addition, possible compatibilities will also be discussed.

Both disciplines agree on setting the study of "habits" and "habit learning" in a global frame of human actions. And here, again, there seems to be a source of incompatibility between philosophers and neuroscientists. As will be discussed later on, one of the main criticisms to human free will from science has come from the experiments carried out by Benjamin Libet. For some investigators, philosophers and experimental researchers alike, Libet's paradigm demonstrates that the conscious decision to act is a mere illusion after the brain activity has already initiated the action. Then, how could habits (in the classical philosophical understanding of the term) be the essence of free will when a single human action could lack of a conscious decision?

In this chapter, we shall try to uncover free will in human actions from a twofold approach: first, we will discuss whether Libet's experiments posit a real threat to free will in single human actions; second, we will analyze how neuroscience has studied habit learning over the years, and will discuss the limitations emerging from it. Doing so we intend to demonstrate that human habits, different from habit learning in animal studies, could be accepted as a proof of free will, instead of a threat to it.

12.2 Libet's Experiments: A Challenge to Free Will?

Free will has been dissociated from human actions from different perspectives (Smith 2011). For instance, Patricia Churchland has claimed that the brain is a causal machine, the only cause of our behaviour, and therefore free will has to be excluded from any scientific analysis of human actions (Churchland 2006). As it has been criticized by Murillo and Giménez-Amaya, this point of view lays in an unproven postulate that avoids a possible rational discussion: if someone says that

free will does not exist, the only proof that can be presented to demonstrate its existence is rejected (Murillo and Giménez-Amaya 2008; in this context see also Giménez-Amaya and Murillo 2009; Giménez-Amaya 2011). Thus, the crucial question that should be answered is as follows: can science prove that free will is just an illusion? According to Blackmore, Benjamin Libet found a way to do it (Blackmore 2007). As a first goal of our note, we would like to spend some time discussing the experiments that have been used to argue about the dissociation of free will from human acts. We will also make a comment on more recent works that seem to confirm Libet's results.

Benjamin Libet (1916–2007) was a researcher at the Department of Physiology of the University of California in San Francisco. He was first interested in the sensation thresholds, that is, the degree of brain stimulation that was needed to artificially provoke a somatic stimulus. This led him to the field of consciousness. Libet decided to design his famous experiment after the discovery of the "readiness potential" (RP) by Kornhuber and Deecke (1965). The RP is a change in electrical activity of certain brain regions preceding the execution of a determined action. Libet wanted to investigate the relationship between RP, the conscious decision to act, and the execution of the movement. Thus, he needed to record the exact time at which these events happened when the human volunteers decided to move their hands. To do so, he designed a particularly fast-paced clock, with a dot that circumscribed the clock face in 2.6 s. Subjects were asked to record where the dot was located when they felt the conscious desire to make the movement. With a much simpler additional experiment, the researchers concluded that the subjects' error when reporting the time was only about 50 ms, so the method could provide meaningful results. In addition to this, the brain activity of the volunteers was recorded by placing some electrodes on the scalp corresponding to motor and premotor areas in the frontal lobe of the brain, responsible for the movement of the hand, whose onset was also precisely measured by an electromyogram. Libet found that, for a well-planned movement, the conscious desire to act appeared about 200 ms before the movement initiation. Surprisingly, the RP started about 550 ms in that condition. To put it in plain words, the brain had already started the movement before the subjects were conscious that they wanted to move their hand. It could be interpreted as if the conscious will to make the movement would be a *consequence* of the brain activity that was preparing for it (Libet et al. 1983).

These findings were very much celebrated by those investigators claiming that free will was incompatible with neuroscience (see Blackmore's comment just mentioned above). However, Libet was more cautious about the interpretation of his experimental results. He said that "the assumption that a deterministic nature of the physically observable world (to the extent that may be true) can account for subjective conscious functions and events is a speculative belief, not a scientifically proven proposition" (Libet 1999). In the same publication, he further clarified that his experiments do not exclude free will, but do illustrate that the volitional process of an action's initiation is unconscious. According to this author, a person can voluntarily restrain from performing an action before starting it, irrespective to the RP; and that would be an exercise of free will.

After Libet, different researchers have tried to investigate whether the RP is really the origin of the awareness of the movement initiation, or the latter is due to an unknown additional phenomenon. Haggard and Eimer, for example, cast doubts on the fact that the RP was analogous to the movement, or on the contrary it was related to other unspecific processes, such as arousal. To investigate this, they studied the lateralized RP (LRP), which is a measure of the difference between RP in both sides of the brain. This is more specific to the movement than the RP alone. Thus, if the right hand is moving, RP should be located on the left hemisphere of the brain; LRP is the subtraction of the RP on the right side of the brain to that registered on the left side (Coles 1989; Eimer 1998). Haggard and Eimer also added certain degree of "freedom" to the paradigm by allowing the subjects to decide to move either the right or the left finger. They proved that LRP onset co-varied with the early or late awareness of the initiation of an action, but RP did not. Therefore, they conclude that "only the smaller temporal discrepancy between LRP onset and [a voluntary movement] awareness may need to be explained by those who wish to retain the traditional concept of free will" (Haggard and Eimer 1999).

However, for the sake of the argument, the issue remains unchanged: an influential stream of neuroscientists keep claiming that the volition to act emerges from the activation of the cerebral cortex that is preparing the movement.

One final example in this group of experiments is the one carried out by Soon and collaborators (Soon et al. 2008). In this occasion, they tried to confirm and extend Libet's experiments with a different methodology using functional magnetic resonance imaging (fMRI). Volunteers lay inside the scanner and performed a task, similar to a very simple computer game. A chain of letters that consecutively appeared on the screen, in a random order, substituted the clock that Libet designed for his paradigm. As in the experiments carried out by Haggard and collaborators, they were instructed to press either the left or right button of a console whenever they felt the urge to do so, and to mentally record the letter that was on the screen when they consciously decided which button to press. Two seconds after the action, the screen changed and showed three of the letters that were displayed just before pressing the button, plus a hash symbol in case the conscious decision was made earlier. Subjects had to indicate which one they were viewing when they made the decision to press the left or right button. Each individual letter remained on the screen for 500 ms. The authors decided to exclude those trials where subjects selected the hash symbol—less than 2% of the total—since they decided to discard those cases in which the conscious decision was made more than 1.5 s before the action.

The results of these experiments were even more surprising than Libet's. Researchers found a specific activation 10 s before the conscious decision of which button to press. However, according to their words, they do not contradict Libet's findings but extend them, because they found this "long-term determinant" of the decision in the frontal pole and the parietal cortex. They did find significant "neural" activity in the motor and premotor cortices, much as in the previous experiments, but this happened closer in time to the actual movement. Therefore, Soon and colleagues

conclude that the activity of the brain had selected the action even 10 s before the subjects were aware of the decision. Although all these experiments will be further commented on, it is worth noting that fMRI provides an *indirect* measure of neural activity, since it indicates the blood supply that the brain is receiving in the context of a task. This measure is delayed with respect to the activity of neurons, and therefore it is said that fMRI has low temporal resolution.

As a summary of this first descriptive section of our paper, the following points should be kept in mind: (1) Libet found that the cortical activity to trigger a movement starts before experiencing the urge to act; (2) Haggard and Eimer localized this activity in the contralateral motor-related cerebral cortex; (3) Soon and collaborators described a peak of activity in the frontal and parietal cortices 10 s before the conscious selection of which hand to move; (4) as a corollary to Libet's and subsequent experiments, several neuroscientists and philosophers have claimed that free will is just an illusion.

12.3 Are Libet's Experiments About Free Will?

At this point, we should re-formulate the main question that underlies this commentary: is Neuroscience compatible with free will? We have discussed how the approach of this discipline to human acts points to a negative answer, since simple motor actions seem to be determined by one's brain activity before the conscious decision to act. But, is it possible to find a re-interpretation of these scientific facts, contextualizing them in a more holistic vision of the human being?

The first issue to be addressed should be what Libet's results are telling us. In order to obtain meaningful results in neuroscientific research, the conditions of the experiment must be controlled to minimize contaminating brain activity. Even doing so, the researcher should be cautious in the interpretation of the results, due to the high complexity of the human brain. A cardiologist might be sure that certain damage to the heart is going to be manifested by certain symptoms in the patient; however, every neuroscientist knows that, although some generalizations may be helpful, there is not such clear correlation. This is the reason neuroscientific experiments must be kept as simple as possible. Likewise, interpretations should carefully deal with issues that have not been fully proved by the experiments.

In Libet's paradigm, the researcher is analyzing the voluntary movement of the participant's hand. If we try to frame the conditions under which this experiment was carried out, we will probably conclude that such scenario has little to do with free will. The participants knew that their own role was doing what the researchers asked of them. Whether the movement was instructed or self-initiated, they knew that the only thing they had to do was move the hand in a particular time frame. Hence, it is hard to find a direct relationship with the existence of free will.

Haggard and collaborators, as well as Soon and colleagues, tried to address this issue by adding a "free choice" factor, and leaving the participants to select which hand to move. Again, it seems difficult to argue against the existence of free will

taking into account this experimental framework. Although it is beyond the scope of our paper to give a thorough philosophical account of free will, it should be considered that a "free action" is not just a cause, linked to consciousness, which is able to modify the physical world (Murillo and Giménez-Amaya 2008). According to these authors, if we understand free will from an Aristotelian point of view, it arises from the *proaíresis*—usually translated as "choice". This is something exclusive to human beings and it is not manifested as a causal agent, but as making a choice in the context of one's own life experience.

But this complex view of human freedom is found not only in researchers with an Aristotelian philosophical background. In a recent work, Patrick Haggard acknowledges that "in humans, the rules mapping stimuli and contexts to appropriate responses can become extremely complex" (Haggard 2011). Therefore, he seems to separate Libet's and his own experiments from free will in real life—although later on in the same paper he expects that neuroscience will be able to demonstrate eventually its source in the human brain.

Bennett and Hacker, in their well-known text "Philosophical foundations of Neuroscience", also argued against the pretended analogy between Libet's paradigm and free will (Bennett and Hacker 2003). They are mainly addressing neuroscientists and reminding them that volition and will-power have a wide range of concepts, such as felt inclinations, felt desires, wanting, purpose, goal and aim, decision, intention and so on. Besides, they present different categories of a volitional act: voluntary, involuntary or non-voluntary, intentional or unintentional, deliberate and impulsive, etc. Their goal is to demonstrate that a voluntary human action is not just a sort of behaviour caused by acts of will, volitions, wants, intentions or decisions.

To put some order among all these elements, they first distinguish acts (things that one can do or abstain from doing), from doing things that are not acts (such as falling sleep). Human acts can be voluntary or not. The former can be classified in intentional, unintentional and non-intentional, whereas the latter can be involuntary, non-voluntary and under duress. In the context of free will, voluntary acts are the most important for our purpose. Bennet and Hacker admit that there is not a *feeling* that tells us some act we have performed has been voluntary; we just *know* it. In addition, a voluntary movement must be voluntary in its inception, continuation and termination. For that reason, for example, blinking is only partially voluntary. The most striking statement made by these authors, one might think, is that a voluntary movement is not a movement caused by an inner act of volition. In the authors' opinion, will-power is not a mental equivalent of muscle power, but determination and persistence in pursuit of one's goals in the face of difficulties. That is, something very close to the Aristotelian concept of *proaíresis* mentioned above.

Furthermore, they offered a careful explanation of why a volitional antecedent is *not* the cause of an act. This explanation can be summarized as follows: if it were so, the mere act of intending or wanting to do something, would be enough to have that act done. In other words, after deciding I want to drink water from a bottle I could just relax and let the action be performed by itself. Following this, it is impractical to think that any voluntary act should be preceded by a volitional cause.

As Bennet and Hacker exemplify, that would mean that writing each letter of a word, and each word of a sentence, should be caused by a different volitional event (Bennett and Hacker 2003).

Then, what happens with Libet's experiments? First of all, Bennet and Hacker think that Libet's paradigm is based on a confused presupposition: an act is voluntary as long as it is preceded by a feeling of desiring, wishing, wanting or intending to perform it. That is, feelings of volition are not necessary, neither sufficient, for voluntary movement. They give a clear example of this: feeling the urge to sneeze right before doing so does not transform it into a voluntary act. In fact, they stated that a movement that is caused by a felt urge is not voluntary. One can feel the urge to do something, but this does not have a univocal causal relation with the act to be done.

In addition, from a methodological point of view, they criticize the fact that volunteers had to report "a feeling of intention to move a hand", when moving one's hand voluntarily does not normally involve such feeling. These authors end their critical reading of Libet's experiments by reminding us that many of our acts are planned or decided in advance, even days or months before being done. Then, when the date is due, we do not act feeling urged to do it, but just to fulfil a plan we had designed. If we want to place free will in any step of this picture, it is more likely to be in the time of the decision—perhaps months before the act—rather than in the moment we get up from our couch and dress to fulfil our plan.

It is obvious that Libet's paradigm cannot account for such a wide and scientifically unspecific notion of human freedom. Besides, it should be also admitted then that these experiments by themselves are not a thread for free will from a holistic perspective. At any event, Libet's experiment is a cleverly designed and valuable paradigm to demonstrate how certain parts of the brain are active in preparing a simple and well-known movement.

12.4 Brain Activity and the Problem of Time in Libet's Experiments

Another conceptual and methodological problem that may arise from these experiments is time (Murillo and Giménez-Amaya 2008). Briefly, it can be formulated as follows: how can we compare the timing of the neural processes occurring in the brain with consciousness and with "external" time? The performance of a voluntary movement involves a high number of brain regions, functionally linked by intricate connections (see, for example, Nieuwenhuis et al. 2008). Besides, we could also add to the picture those circuits that are *inhibiting* the movements that might interfere with the one selected to be carried forward. The whole process is hierarchically organized, being the associative cortices situated at the highest level. Then, directly or through non-cortical structures such as the basal ganglia or the cerebellum, the neural information flows to the premotor

cortex, where the movement is planned, and finally to the motor cortex, which sends the final order to the brain stem and the spinal cord (Allen and Tsukahara 1974; DeLong et al. 1986). Although this flow of information has been studied in depth, it is still not known how the different brain areas are synchronized from an *external*—and, hence, measureable—point of view.

We do know that there is a time lag between the moment when a subject is prompted to act, and the actual performance of the action. This delay is known as the "reaction time", and nowadays is a very informative variable in normal and pathological mental states, such as schizophrenia (see, for example, Murray et al. 2008). By means of electroencephalography and sophisticated statistical analysis, Blinowska and collaborators have situated the time difference between the activation of association and premotor cortices in a cognitive task at about 0.7 s, but they have reported it to be highly variable between subjects (Blinowska et al. 2010).

Even considering that the exact timing of brain activity and action can be measured, how can we precisely assess the relationship of the former with a *conscious decision*? This is a potential source of problems in Libet's paradigm. As mentioned above, he has tried to overcome it by using a control stimulus, in which subjects had to report the exact time when they were feeling a touch in their hands. The error—that is, the difference between the subject's report and the actual touch—was only 50 ms on average. From this result, Libet inferred that the error in his experimental paradigm had to be the same.

However, can we be sure that the time of a reported sensation is comparable to that of a reported conscious decision? The main problem here is the accuracy with which we consider ourselves conscious of having made a decision. Is it when we have internally verbalized the desire to act? Do we have an internal trigger that initiates the conscious action, analogous to the one that according to Libet triggers the unconscious process? Again, we should admit plainly that there are too many loose ends that should be tied down before drawing all-or-nothing conclusions. In this context, Gallagher has written "that this problem can be solved as long as we do not think of free will as a momentary act. Once we understand that deliberation and decision are processes that are spread out over time, even, in some cases, very short amounts of time, then there is plenty of room for conscious components that are more than accessories after the fact" (Gallagher 1998). One example to such elements, Gallagher claims, could be feedback loops, whose importance has been extensively proven in relation to consciousness (Raffone and Pantani 2010).

Then, if the neural activity that Libet, Haggard and Soon have found is not causally related with the conscious volitional initiation of the action, what could it mean? One possible explanation has been already outlined: the premotor cortex is actually getting ready to perform the movement through the motor cortex. If that is so, Libet's interpretation at this instance seems correct.

But Zhu, in his work "Reclaiming volition" (Zhu 2003), has tried to find an alternative explanation to Libet's experiments. First of all, he highlights two possible flaws in the interpretation of the results. The first is similar to the one mentioned by Bennett and Hacker, arguing that the conscious decision to act is not taken by the subject in each trial, but at the beginning of the experiment when they

decided to take part in it. Secondly, Zhu thinks that the origin of the "urge to move" is not an unconscious activation in the motor cortex, but the actual instruction of monitoring internal processes, which otherwise should have been unconscious—as Bennett and Hacker pointed out, one usually does not feel the "urge to move" when grabbing a glass of water. To prove this second misinterpretation, Zhu cites the work by Keller and Heckhausen (1990), who demonstrated that the RP was present in both conscious reported and non-reported movements. Zhu finishes his interpretation with these clear words: "Therefore the functional role of volition in initiating voluntary actions is not undermined by Libet's experimental studies. We are not only the censor or controller, but also the author or originator of our own actions. This common-sense image of human agency, which is fundamental to our understanding of responsibility, freedom and human dignity, can be preserved".

As a final note on the alternative meanings of the activations found in these experiments, it should be taken into account that an activation of the premotor or motor cortex is *not always* followed by a movement. There are several intriguing examples of this that have been recently found.

The first comes from the mirror neuron system recently discovered by Rizzolatti and Craighero (2004). These neurons are activated when a subject performs an action, and also when that action is just seen. Rizzolatti and collaborators found them in a monkey having a piece of food, where they found that the same neurons were active when the piece of food was taken by another monkey or by a human (Rizzolatti et al. 1996). The presence of a putative mirror neuron system in humans have been also suggested, both with functional magnetic resonance imaging (Gazzola & Keysers 2009) and individual neuronal recordings in epileptics (Mukamel et al. 2010). In the context of our paper, the interesting part of these results is the activation of motor and premotor areas (among others) even though there is no movement at all.

The second example at this respect is motor imagery, that is, the "mental rehearsal" of an action without carrying it out. Recently, it has been wonderfully employed as an experimental tool by Adrian Owen to demonstrate volitional brain activity in patients in a minimally conscious state (close to coma). These patients are obviously not able to communicate with others, and therefore it cannot be demonstrated that they are able to listen to or understand instructions from others. However, Owen designed an experiment to do this by using mental imagery. Firstly, his team found a consistent activation in the brain of healthy volunteers after motor imagery (Boly et al. 2007). This happened even in complex tasks such as spatial navigation, involving a wide network of brain regions that are activated in actual navigation. The only areas that showed a common activation in all tasks were the premotor cortex and the pre-supplementary motor cortex, both target regions in Libet and Haggard experiments. Secondly, they found the same brain activity in a patient in a vegetative state (Owen et al. 2006). In addition to the scientific and ethical implications of these findings, they clearly prove that it is possible to find a strong activation in motor and premotor cortices unrelated with an eventual movement. In the same line of thought, Kilner and collaborators

demonstrated that, in humans, an activation of the motor cortex is also triggered when anticipating a predicted movement performed by others (Kilner et al. 2004).

One of the main criticisms to Libet's work, admitted by him, was that they did not record the neural activity in those trials in which subjects felt the "urge to move" but ultimately did not do it. Some authors have argued that, should there be the same RP in those trials, it would not condition the final decision of the subject to either move or not to move (Murillo and Giménez-Amaya 2008). On the whole, the examples mentioned above—mirror neurons and motor imagery—show that activation of premotor and motor cortices are not sufficient to make an actual movement.

12.5 Human Acts and Free Will

As a summary of our intention in the first section of this note, we have intended to provide enough evidence to demonstrate that Libet's experiments, as well as more recent ones, are not such a threat to free will in human actions as was thought by some philosophers and neuroscientists.

The next section aims to go one step further in exploring human actions, and we would like to analyze the case of habit learning. From a neuroscientific standpoint, habit learning has been restricted to a narrow view of the phenomenon, that is, the development and consolidation of motor automatisms or routines. In the next part of our paper we will first describe habit learning from the perspective of neuroscience, and finally we will discuss the possible convergence of this view with other philosophical outlooks.

12.6 Habit Learning in Neuroscience

The history of habit learning research in neuroscience is relatively short, as has been recently explained by Seger and Spiering (2011). According to these authors, the term "habit" was first used in modern psychology by William James (1890), and researchers on animal learning exploited the term to define a motor routine that was triggered after certain stimuli. This view has been very much preserved until the present time. During the twentieth century, the concept of habit learning has gained new features, thanks to the contribution from cognitive experimental psychologists.

Thus, in 1957 Scoville and Milner reported that memory depended on the integrity of the medial temporal cortex and, in particular, the hippocampus (Scoville and Milner 1957). These results were confirmed in patients with amnesia that showed hippocampal damage, and experimental research in hippocampus-ablated animals demonstrated that these animals were unable to remember things previously learned. However, Hirsh (1974) found that learning of new routines—and, therefore, some sort of memory-related events—was possible in animals with hippocampal lesions. This and subsequent findings led to a re-evaluation of memory

processes in the brain, distinguishing two types of memory: "declarative" or "explicit", and "non-declarative" or "implicit". The former was related to the hippocampus and the latter, with an unknown biological substrate, included the so-called "habit learning". Squire and Zola-Morgan, who further clarified this classification, pointed to the striatum—part of the basal ganglia, a group of brain nuclei located under the cerebral cortex—as a putative substrate for habit learning (Squire and Zola-Morgan 1991). Nowadays, the role of the striatum is widely accepted, together with its prominent neural connectivity with the cerebral cortex, as the neurobiological substrate of this process (Graybiel 2008).

Going back to the early psychological research on habit learning, it is interesting to analyze what "implicit" or "non-declarative" learning implies. First, it refers to an *unconscious* process. Thus, whereas learning a fact through declarative memory involves certain degree of awareness in its acquisition or retrieval, subjects cannot verbally explain what they have learned in implicit learning (Seger 1994). This has been extensively demonstrated in psychological studies of serial reaction time or artificial grammar tasks (Squire and Zola 1996), where subjects successfully perform a task unaware of its strategy. Another feature of this kind of learning is its *automaticity*, which is uncovered by the dual task experiment: a non-declarative learned task can be carried out simultaneously with a more attention-demanding novel task. This also supports the unconscious aspect of habit learning.

There have been further contributions to the automaticity or *rigidity* of implicit learning from experimental psychologists. Dickinson has distinguished between goal-directed behaviour and habits in an instrumental learning context (Dickinson 1985). The main difference between them is the purpose of the animal to perform an action. Dickinson observed that, after a number of trials, rats learned the contingency between a lever press and delivery of food. When the task had been repeated a high number of times, rats kept pressing the lever even though they were sated and food was not delivered anymore. Therefore, Dickinson concluded that rats have developed the "habit" of pressing the lever, irrespective to their goal or the outcome of the action. His experiments have been also used to state that habit learning is *slow*, since it needs many repetitions of an action to become a "habit".

During the last two decades, Graybiel has proposed a more refined neurobiological model for habit learning. According to her research, sequences of actions are *chunked* into smaller subunits that have a correlation with the chunked activity of neurons in the prefrontal cortex and striatum (Graybiel 2008). She agrees with Dickinson's proposal in that the first few times an animal performs an action it is goal-directed. When the animal has found the most effective way to proceed—through the dopaminergic system, according to Graybiel—it chunks the action into smaller fragments to optimize it. Then, consciousness and flexibility will only appear at the beginning of each chunk, when the neuronal activity is the highest. The purpose of this neuronal activity is simple: to discharge consciousness from the continuous supervision of an action's performance, and therefore to transform it into a less demanding activity.

Several authors have proposed that in certain psychiatric diseases, such as obsessive-compulsive disorder (OCD) and addiction, there is an exaggerated translation of actions into habits. Patients with OCD present repetitive, ritualistic

behaviours that they keep doing even though they are identified as pointless and unproductive. Gillan and collaborators have recently hypothesized whether in OCD patients there was an imbalance between goal-directed action control and habitual behaviour (Gillan et al. 2011). To do so, they instructed the volunteers to learn the contingencies between the selection of an action and a beneficial outcome. At some point of the experiment, they were informed that certain associations had changed, and some actions would stop being beneficial, and they were asked to choose the rewarding actions instead of the neutral one—outcome devaluation test. Finally, they entered into a "slip-of-action" phase, where instead of choosing between two actions, they had either to perform the action—press a button—or to refrain from it—withhold the button press. Even though OCD patients did not find any problem learning the game, they made significantly more mistakes than the control group in the outcome devaluation test, as well as in the slip-of-action test. This study shows, as has also been suggested for drug addiction (Everitt and Robbins 2005), that in these pathological conditions patients' behaviour is better explained by "habitual responding" instead of being directed by the most valuable outcome. Yin and collaborators, in a study on rats, have demonstrated that the neural basis of this imbalance is located in the striatum (Yin et al. 2004).

In summary, from a neuroscientific perspective, habit learning is viewed as the performance of an action that has been previously learned after many repetitions, in an unconscious manner, and whose execution is inflexible and independent to the outcome. This view could be seen as a putative threat to free will in humans, whose life depends on habits at a wide extent. Habit learning involves a loss of flexibility, evaluation of strategies and, therefore, free will.

12.7 A Note on Habit Learning and Free Will

In the content of this paper, we have occasionally mentioned that the view of habits from neuroscience seems to be focused on habit learning. However, the classical philosophical perspective is much broader. Let us make now a brief philosophical account of this following Aristotle and Aquinas on the subject. For them, a habit is understood as a quality or disposition that has an effect on ourselves and, certainly, has also an impact on others. This relation can be positive or negative and, therefore, a habit can be described as a virtue or a vice, respectively (Polo and Llano 1997). In this context, habits are the base of freedom and are in fact based on freedom. Moreover, they are also the ontological basis of science, art, morality and society (Polo 1996).

As happens in neuroscience, the philosophical account of habit formation is mostly connected to actions, as much as humans express themselves by *acting*. Habits are perfections acquired by action. In other words, both Aristotelian-Thomistic philosophical tradition and neuroscience agree in that the repetition of an action, leading to an improved performance, is the basis of habits and habit learning in sensorimotor and emotional contexts. Again, the main problem seems to stem from what each discipline considers as a human action and, more importantly,

from the way in which the holistic view of a person is taken into account, and not restricted to the study of a particular and isolated action.

Considering this, the description of habits from the Aristotelian-Thomistic philosophical tradition and habit learning research in neuroscience might not be so different after all. For example, consciousness has been reported to be decreased when performing an action that has become a habit. From the philosophical standpoint, this could not only be acceptable, but also desirable: the lack of a continuous conscious supervision of certain types of action could explain the appearance of a *tendency* to act in a particular way. Thus, doing a good action for the first time might require the complete engagement of consciousness during the whole process. When the same or different good actions have been repeated over time, the individual will tend to act well and justly in a prompt and easy way. This is one of the key points in the view of habit formation by the Aristotelian-Thomistic philosophical tradition, and we find it not at all incompatible with how neuroscience understands the process of habit learning in the sensorimotor and emotional contexts of our behaviour. In fact, we could find some analogies in both types of acquisition for the human actions to be performed.

One might think that the main difference between both accounts appears precisely when freedom—or free will—enters into the subject. Thus, from a neuroscientific perspective it could seem that if an individual was acting only based on habits, he or she would become an automaton. Such could be seen in the case of the OCD research mentioned above, as well as in those carried out in animal models that develop stereotypes, Tourette's syndrome patients, and so on.

This apparent controversy might be addressed by considering again the human being as a person, and not just as the subject of a particular and isolated action. As an example, this can be exemplified by an OCD patient who needs to wash his hands many times before going out—a quite common symptom in this disease (Rettew et al. 1992). It is obvious that personal hygiene is a positive habit. However, the fact that a patient exhibits compulsive hand washing does not mean he cares about his hygiene as a whole. In fact, as often the case, it is likely that this patient could neglect his overall cleanliness by compulsively focusing on his hands. These altered and imbalanced motor routines are somehow "superimposed" in a person with a positive habit of hygiene. In our opinion, this is precisely the symptom that shows an alteration of the nervous system: the dominance of sensorimotor or emotional routines over habits (as understood in the Aristotelian-Thomistic philosophical tradition). Routines and habits are present both in the healthy subject and the OCD patient but, whereas in the former there is a concomitant performance, the abnormal routines cast a shadow over habits in the latter.

Through the neuroscientific study of habit learning, it has been said that an action is improved at the same time that it becomes inflexible. In the context of human actions, this could be seen as a new proof of confrontation between habit learning and free will. Inflexibility is undoubtedly true, again, for a single action, but arguable in the context of the set of actions that configure behaviour. Let's think for example of a basketball player who practices every day to improve certain sensorimotor routines. If he only mastered a particular move, his game would be

predictable and it would be quite easy for his opponents to defend against him. However, as much as he acquires a set of routines, the flexibility of his gameplay will increase by deciding which one to perform according to the opponent.

Finally, we think that there is a fact that should be taken into account: when a person improves the performance of a particular action, it will also have a positive impact on other related actions. If this were not true, every single human action should be independently practiced to be mastered. We will conclude this section of our paper with a final example to clarify this. An amateur piano player is recruited by a scientist for a "habit learning" experiment. She is required to learn a particular sequence of finger tapping while following the instructions in a computer screen. When her data are analyzed, the scientist corroborates that this piano player improves her performance in that particular sequence of movements, since she does it quickly every trial. However, when a new finger move is unexpectedly included in the sequence, the piano player usually fails to accommodate it, demonstrating the inflexibility of the motor routine learned. When she gets home, the amateur musician spends 1 h a day playing simple scales on the piano to improve her skills. After several months, the scientist publishes a paper showing that "habit learning" improves performance at the cost of flexibility. However, the piano player has the experience that she can play the well-practiced scales much faster than before, and also that she needs to pay less attention to how her fingers are moving and she can start improvising. She is now a better piano player because, through certain motor routines, her performance is more spontaneous, much more flexible.

12.8 Conclusion

This chapter tries to show that neurobiological research is far from being incompatible with free will. Our main goal was to discuss the different standpoints that neuroscience and philosophy (according to the Aristotelian-Thomistic tradition) have on habit learning, and their relationship with free will. Due to the fact that habit learning is intrinsically related to human actions, we have first commented on Libet's and others' experiments that seem to explain free will in human acts as an illusion, since brain activity is in command of our actions. We have tried to provide evidence to illustrate how free will should be put in a wider context, thus showing it is not refuted by neuroscientific experiments.

Having tackled this problem, we have tried to clarify whether habit learning within a broader philosophical background supports free will, since human actions are observed in a more holistic way. We have pointed out that habits should not be misunderstood as sensorimotor and emotional routines, but as qualities or dispositions through which human behaviour is improved—or, in the case of vices, degenerates. We have tried to provide some examples to illustrate this point, and have suggested that certain psychiatric illnesses, such as OCD, should not be considered as an increase in "habitual behaviour", but as an imbalance and

dissociation between the development of sensorimotor and emotional routines, and the person as a whole configured by habits.

References

Allen, G. I., & Tsukahara, N. (1974). Cerebrocerebellar communication systems. *Physiological Reviews, 54*(4), 957–1006.

Bennett, M., & Hacker, P. (2003). *Philosophical foundations of neuroscience*. Oxford: Blackwell Publishing.

Blackmore, S. (2007). Mind over matter? Many philosophers and scientists have argued that free will is an illusion. Unlike all of them, Benjamin Libet found a way to test it. *The Guardian Unlimited*. Retrieved from http://commentisfree.guardian.co.uk/sue_blackmore/.

Blinowska, K., Kus, R., Kaminski, M., & Janiszewska, J. (2010). Transmission of brain activity during cognitive task. *Brain Topography, 23*(2), 205–213.

Boly, M., Coleman, M. R., Davis, M. H., Hampshire, A., Bor, D., Moonen, G., et al. (2007). When thoughts become action: An fMRI paradigm to study volitional brain activity in non-communicative brain injured patients. *NeuroImage, 36*(3), 979–992. doi:10.1016/j.neuroimage.2007.02.047.

Churchland, P. (2006). *Patricia and Paul Churchland in Conversations on consciousness: What the best minds think about the brain, free will and what it means to be human*. (S Blackmore, Ed.). New York: Oxford University Press.

Coles, M. G. (1989). Modern mind-brain reading: Psychophysiology, physiology, and cognition. *Psychophysiology, 26*(3), 251–269.

DeLong, M. R., Alexander, G. E., Mitchell, S. J., & Richardson, R. T. (1986). The contribution of basal ganglia to limb control. *Progress in Brain Research, 64*, 161–174. doi:10.1016/S0079-6123(08)63411-1.

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society B: Biological Sciences, 308*(1135), 67–78. doi:10.1098/rstb.1985.0010.

Eimer, M. (1998). The lateralized readiness potential as an on-line measure of central response activation processes. *Behavior Research Methods, 30*, 146–156.

Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: From actions to habits to compulsion. *Nature Neuroscience, 8*(11), 1481–1489. doi:10.1038/nn1579.

Gallagher, S. (1998). The neuronal platonist. *Journal of Consciousness Studies, 5*, 706–717.

Gazzola, V., & Keysers, C. (2009). The observation and execution of actions share motor and somatosensory voxels in all tested subjects: Single-subject analyses of unsmoothed fMRI data. *Cerebral Cortex (New York, N.Y.□: 1991), 19*(6), 1239–1255. doi:10.1093/cercor/bhn181.

Gillan, C. M., Papmeyer, M., Morein-Zamir, S., Sahakian, B. J., Fineberg, N. A., Robbins, T. W., et al. (2011). Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder. *The American Journal of Psychiatry, 168*(7), 718–726. http://ajp.psychiatryonline.org/article.aspx?articleid=115977.

Giménez-Amaya, J. M. (2011). A better understanding of freedom. An interdisciplinary approach to neuroscience and philosophy. In J. Sanguineti, A. Acerbi, & J. Lombo (Eds.), *Moral behavior and free will: A neurobiological and philosophical approach* (pp. 47–60). Morolo: IF Press.

Giménez-Amaya, J. M., & Murillo, J. I. (2009). Neurociencia y libertad: Una aproximación interdisciplinar. *Scripta Theologica, 41*, 13–46.

Graybiel, A. M. (2008). Habits, rituals, and the evaluative brain. *Annual Review of Neuroscience, 31*, 359–387. doi:10.1146/annurev.neuro.29.051605.112851.

Haggard, P. (2011). Does brain science change our view of free will? In R. Swinburne (Ed.), *Free will and modern science* (pp. 7–24). New York: Oxford University Press.

Haggard, P., & Eimer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. *Experimental Brain Research, 126*(1), 128–133.

Hirsh, R. (1974). The hippocampus and contextual retrieval of information from memory: a theory. *Behavioral Biology, 12*(4), 421–444.

James, W. (1890). *Principles of psychology*. New York: Henry Holt.

Keller, I., & Heckhausen, H. (1990). Readiness potentials preceding spontaneous motor acts: Voluntary vs. involuntary control. *Electroencephalography and Clinica Neurophysiology, 76*(4), 351–361.

Kilner, J. M., Vargas, C., Duval, S., Blakemore, S.-J., & Sirigu, A. (2004). Motor activation prior to observation of a predicted movement. *Nature Neuroscience, 7*(12), 1299–1301. doi:10.1038/nn1355.

Kornhuber, H. H., & Deecke, L. (1965). Changes in the brain potential in voluntary movements and passive movements in man: Readiness potential and reafferent potentials. *Pflüger Archiv für die Gesamte Physiologie des Menschen und der Tiere, 284*, 1.

Libet, B. (1999). Do we have free will? *Journal of Consciousness Studies, 8*, 47–57.

Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). *Brain, 106*, 623–642.

Mukamel, R., Ekstrom, A. D., Kaplan, J., Iacoboni, M., & Fried, I. (2010). Single-neuron responses in humans during execution and observation of actions. *Current Biology: CB, 20*(8), 750–756. doi:10.1016/j.cub.2010.02.045.

Murillo, J. I., & Giménez-Amaya, J. M. (2008). Tiempo, conciencia y libertad: consideraciones en torno a los experimentos de B. Libet y colaboradores. *Acta Philosophica, 11*(17), 291–306.

Murray, G. K., Clark, L., Corlett, P. R., Blackwell, A. D., Cools, R., Jones, P. B., et al. (2008). Incentive motivation in first-episode psychosis: A behavioural study. *BMC Psychiatry, 8*(8), 34. doi:10.1186/1471-244X-8-34.

Nieuwenhuis, S., Voogd, J., & Van Huijzen, C. (2008). *The human central nervous system*. Heidelberg: Springer.

Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., & Pickard, J. D. (2006). Detecting awareness in the vegetative state. *Science (New York, N.Y.), 313*(5792), 1402. doi:10.1126/science.1130197.

Polo, L. (1996). Tener y dar. *Sobre la existencia cristiana* (pp. 103–135). Pamplona: EUNSA.

Polo, L., & Llano, C. (1997). Los hábitos. *Antropología de la acción directiva* (pp. 103–112). Madrid: Unidad Editorial.

Raffone, A., & Pantani, M. (2010). A global workspace model for phenomenal and access consciousness. *Consciousness and Cognition, 19*(2), 580–596. doi:10.1016/j.concog.2010.03.013.

Rettew, D. C., Swedo, S. E., Leonard, H. L., Lenane, M. C., & Rapoport, J. L. (1992). Obsessions and compulsions across time in 79 children and adolescents with obsessive-compulsive disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, 31*(6), 1050–1056.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27*, 169–192. doi:10.1146/annurev.neuro.27.070203.144230.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Brain Research. Cognitive Brain Research, 3*(2), 131–141.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry, 20*(1), 11–21.

Seger, C. A. (1994). Implicit learning. *Psychological Bulletin, 115*(2), 163–196.

Seger, C. A., & Spiering, B. J. (2011). A critical review of habit learning and the basal ganglia. *Frontiers in Systems Neuroscience, 5*, 66. doi:10.3389/fnsys.2011.00066.

Smith, K. (2011). Neuroscience vs philosophy: Taking aim at free will. *Nature, 477*(7362), 23–25.

Soon, C. S., Brass, M., Heinze, H.-J., & Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience, 11*(5), 543–545. doi:10.1038/nn.2112.

Squire, L. R., & Zola, S. M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences of the United States of America, 93*(24), 13515–13522.

Squire, L. R., & Zola-Morgan, S. (1991). The medial temporal lobe memory system. *Science, 253* (5026), 1380–1386.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience, 19*(1), 181–189. http://onlinelibrary.wiley.com/doi/10.1111/j.1460-9568.2004.03095.x/abstract.

Zhu, J. (2003). Reclaiming volition. *Journal of Consciousness Studies, 19*(11), 61–77.

Chapter 13
Free Will and Neuroscience: Revisiting Libet's Studies

Alfred R. Mele

Abstract Benjamin Libet contends both that "the brain 'decides' to initiate or, at least, prepare to initiate [certain actions] before there is any reportable subjective awareness that such a decision has taken place" and that "If the 'act now' process is initiated unconsciously, then conscious free will is not doing it." Elsewhere, I have argued that the claims I just reported are not justified by the data Libet and others offer in support of them. Here I review some of the problems one encounters in attempting to move from Libet's data to his conclusions.

Keywords Libet's experiments • Free will • Awareness • Readiness potential • Proximal decisions and intentions • Distal decision and intentions • Conscious decisions

13.1 Introduction

Benjamin Libet contends both that "the brain 'decides' to initiate or, at least, prepare to initiate [certain actions] before there is any reportable subjective awareness that such a decision has taken place" (Libet, 1985, p. 536)[1] and that "If the 'act now' process is initiated unconsciously, then conscious free will is not doing it" (Libet, 2001, p. 62; see 2004, p. 136).[2] He also claims that once we become conscious of our decisions, we can exercise free will in vetoing them (1985, 1999, 2004, pp. 137–149). Some people follow Libet part of the way: they accept

[1] Elsewhere, Libet writes: "the brain has begun the specific preparatory processes for the voluntary act well before the subject is even aware of any wish or intention to act" (1992, p. 263).

[2] For a useful discussion of what the initiation of an action might amount to and of connections among action initiation, Libet's data, and free will, see Bayne (2011).

A.R. Mele (✉)
Florida State University, Tallahassee, FL, USA
e-mail: amele@fsu.edu

A. Suarez and P. Adams (eds.), *Is Science Compatible with Free Will?: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience*,
DOI 10.1007/978-1-4614-5212-6_13,

his claims about when and how decisions to act are made but reject the window of opportunity for free will as illusory (Hallett, 2007; Wegner, 2002, p. 55).

Elsewhere, I have argued that the claims I just reported are not justified by the data Libet and others offer in support of them (Mele, 2009). Here I review some of the problems one encounters in attempting to move from Libet's data to his conclusions.

13.2 Libet's Studies

Libet's findings are based on a creative series of studies (for summaries, see Libet, 1985, 2004). In some of the studies, subjects are regularly encouraged to flex a wrist whenever they wish. In subjects who do not report any advance planning of their movements, electrical readings from the scalp (EEGs)—averaged over at least 40 flexes for each subject—show a shift in readiness potentials (RPs) beginning about 550 milliseconds (ms) before the time at which an electromyogram (EMG) shows relevant muscular activity to begin. These are type II RPs. Subjects who are not regularly encouraged to aim for spontaneity or who report some advance planning produce RPs that begin about half a second earlier—type I RPs. The same is true of subjects instructed to flex at a prearranged time (Libet, Wright, & Gleason, 1982, p. 325). (According to a common use of "readiness potential," it is a measure of activity in the motor cortex that precedes voluntary muscle motion and, by definition, EEGs generated in situations in which there is no muscle motion do not count as RPs. Libet's use of the term is broader. For example, because there is no muscle motion in the veto experiment I describe later, some scientists would not refer to what Libet calls "the 'veto' RP" (1985, p. 538) as an RP.)

Subjects are also instructed to recall where a dot was on a special clock when they first became aware of something, x, that Libet variously describes as a decision, intention, urge, wanting, will, or wish to move. (The dot on this Libet clock makes a complete revolution in less than 3 s.) On average, the onset of type II RPs preceded what subjects reported to be the time of their initial awareness of x (time *W*) by 350 ms. Time *W*, then, preceded the beginning of muscle motion (a muscle burst) by about 200 ms. The results may be represented as follows:

Libet's results for type II RPs		
−550 ms	−200 ms	0 ms
RP onset	Reported time *W*	Muscle begins to move

(Libet finds evidence of what he regards as an error in subjects' recall of the times at which they first become aware of sensations (1985, pp. 531, 534). Correcting for it, time *W* is −150 ms.)

Again, in Libet's view, consciousness opens a tiny window of opportunity for free will in his subjects. If a subject becomes aware of his decision or intention at −150 ms, and if by −50 ms his condition is such that "the act goes to completion

with no possibility of its being stopped by the rest of the cerebral cortex" (Libet, 2004, p. 138), his window is open for 100 ms. Libet writes: "The role of conscious free will [is] not to initiate a voluntary act, but rather to control whether the act takes place. We may view the unconscious initiatives as 'bubbling up' in the brain. The conscious-will then selects which of these initiatives may go forward to an action or which ones to veto and abort" (1999, p. 54).

13.3 Conceptual Background

Some conceptual background will prove useful for the purposes of assessing the implications of Libet's findings. I start with the concept of deciding to do something—practical deciding. (Deciding that something is true is a distinct phenomenon.) Like many philosophers, I take *deciding* to *A* to be an action—as I see it, a momentary action of forming an intention to *A* (Mele, 2003, ch. 9). Deliberating about what to do is not a momentary action, but it must be distinguished from an act of deciding that is based on deliberation.

This conception of practical deciding does not entail that all intentions are formed in acts of deciding. In fact, many intentions seem to be acquired without being so formed (see Mele, 2003, ch. 9). If, as I believe, all decisions about what to do are prompted partly by uncertainty about what to do (2003, ch. 9), in situations in which there is no such uncertainty, no decisions will be made. Even so, intentions may be acquired in these situations.

Some decisions and intentions are about things to do straightaway. They are *proximal* decisions and intentions. Others—*distal* decisions and intentions—are about things to do later. Ann's decision to phone Al now is a proximal decision; her decision to phone Bob tomorrow is a distal decision. Libet's attention to decisions and intentions is focused on the proximal kind.

Deciding to do something should be distinguished from wanting (or having an urge) to do it. Sometimes people want to do things that they decide not to do. And often, when people want to do each of two incompatible things—for example, meet some friends for lunch at noon and attend a lecture at noon—they settle matters by deciding which one to do. Just as deciding should be distinguished from wanting, so should intending. Intending to do something is more tightly connected to action than is merely wanting to do it.

For critiques of alternative accounts of deciding, see Mele, 2003, ch. 9. A virtue of the account just sketched, for the purposes of this article, is that it is consonant with Libet's apparent conception of practical deciding.

13.4 What Happens at −550 ms?

One inference Libet makes on the basis of his findings is that the brain produces a proximal decision or intention to flex about one-third of a second before the subject becomes aware of that decision or intention. Is this inference warranted?

An alternative hypothesis is that what the brain produces around −550 ms is a potential cause of a subsequent proximal decision or intention to flex and the decision or intention emerges significantly later.

How might one get evidence about whether the onset of the type II RPs at −550 ms is correlated with unconscious proximal decisions or intentions to flex or instead with potential causes of decisions or intentions? An apt question to ask in this connection is how long it takes a proximal intention to flex to generate a muscle burst. If, in fact, the brain produces proximal decisions or intentions in Libet's study about 550 ms before the muscle burst, then, in his subjects, it takes those decisions or intentions about 550 ms to produce a muscle burst. Is this a realistic figure?

Some reaction-time studies provide relevant evidence. In a study in which subjects were watching a Libet clock, the mean time between the sounding of the go signal and the muscle burst was 231 ms (Haggard & Magno, 1999, p. 104). Subjects were instructed to respond as rapidly as possible to the go signal by pressing a button. If detection of the go signal produced a proximal intention to press the button, then the mean time between a subject's acquiring a proximal intention to press and the muscle burst was less than 231 ms. (Detecting a go signal takes time.) And notice how close this is to Libet's time *W*—his subjects' reported time of their initial awareness of something he variously describes as an intention, urge, wanting, decision, will, or wish to move (−200 ms). Even without putting much weight on the exact number (−231 ms), one can fairly observe that if proximal intentions to flex are acquired in Libet's studies, the finding just reported makes it look like a much better bet that they are acquired around time *W* than that they are acquired around −550 ms.

Someone might object that in reaction-time studies of the kind described, muscle bursts and actions are not produced by proximal intentions but by something else. It may be claimed, for example, that the combination of subjects' *conditional* intentions to press whenever they detect the go signal together with their detection of it produces muscle bursts and pressings without the assistance of any proximal intentions to press. (A typical conditional intention has this form: "if [or when] x happens, do y.") But if this claim is accepted, a parallel claim about Libet's studies should be taken seriously. The parallel claim is that, in Libet's studies, the muscle bursts and actions are not produced by proximal intentions but by the combination of subjects' conditional intentions to flex whenever they detect a conscious proximal urge to flex together with their detection of such an urge. Someone who makes this claim may hypothesize that the onset of the type II RPs at −550 ms is correlated with a potential cause of a conscious proximal urge to flex. Libet's findings do not contradict this hypothesis.

Even if Libet is wrong in claiming that the brain produces proximal intentions or decisions to flex at about −550 ms, his claim about the 100 ms window of opportunity for free will merits attention. Libet's idea is that free will can only be exercised consciously and, therefore, can only be exercised after his subjects become conscious of proximal intentions, decisions, or urges to flex (and before it is too late to stop what is in place from generating a flex). He contends that free will can be exercised only in vetoing the decision, intention, or urge of which the person has

become conscious. An alternative hypothesis is that Libet's subjects exercise free will in making conscious proximal decisions to flex rather than after they become conscious of such decisions (or intentions or urges). Given that Libet's findings do not justify the inference that proximal decisions to flex are made before the subjects are conscious of any such decision, they do not contradict the present hypothesis.

13.5 What Happens Between −550 and 0 ms?

Libet's findings are sometimes said to support the thesis that conscious intentions and decisions play no role in producing corresponding actions. It is claimed that they are caused by the same brain events that cause actions and that they are not themselves in the causal chain that results in action (Lau, Rogers, & Passingham, 2007; Wegner, 2002, pp. 55, 67–70, 317–318). Sometimes the following assertion is offered in support of the preceding one: Subjects' conscious proximal intentions to flex cannot be among the causes of their flexes because those intentions are caused by unconscious brain events (Pockett, 2006, p. 21; Roediger, Goode, & Zaromb, 2008, p. 208). This assertion is misguided, as attention to the following analogous assertion shows: Burnings of fuses cannot be among the causes of explosions of firecrackers because the burnings are caused by lightings of fuses. Obviously, both the lighting of its fuse and the burning of its fuse are among the causes of a firecracker's explosion in normal scenarios. Other things being equal, if the fuse had not been lit—or if the lit fuse had stopped burning early—there would have been no explosion. There is no reason to believe that the more proximal causes of firecracker explosions cannot themselves have causes. Analogously, there is no reason to believe that items that are among the relatively proximal causes of flexes cannot themselves have causes and cannot be caused by unconscious brain events.

Is the brain activity registered by, for example, the first 300 ms of type II RPs—*type 300 activity*, for short—as tightly connected to subsequent flexes as lightings of firecracker fuses are to exploding firecrackers? In fact, no one knows. In the experiments that yield Libet's type II RPs, it is the muscle burst that triggers a computer to make a record of the preceding brain activity. In the absence of a muscle burst, there is no record of that activity. So, for all anyone knows, there were many occasions on which type 300 activity occurred in Libet's subjects and there were no associated flexes.

Libet mentions that some subjects encouraged to flex spontaneously report that they sometimes suppressed conscious proximal urges to flex (1985, p. 538). As he points out, because there was no muscle activation, there was no trigger to initiate the computer's recording of any RP that may have preceded the veto (2004, p. 141). So, for all anyone knows, type 300 activity was present before the urges were suppressed.

It is the urges that these subjects are said to report and suppress. Might it be that type 300 activity is a potential cause of conscious urges to flex in Libet's subjects

and that some subjects make no decision about when to flex—unconsciously or otherwise—until after the conscious urge emerges? And might it be that prior to the emergence of the conscious urge, subjects have no proximal intention to flex? That our urges often are generated by processes of which we are not conscious is not surprising. And if we sometimes make effective decisions about whether or not to act on a conscious urge, so much the better for free will. Moreover, as I have explained, Libet's data do not show that subjects have unconscious proximal intentions to flex before they have conscious ones. The data do not contradict the hypothesis that what precedes these conscious proximal intentions is a causal process that includes no unconscious proximal decisions or intentions to flex.

Libet offers two kinds of evidence to support his claim that subjects have time to veto proximal conscious urges to flex. One kind has already been mentioned: subjects say they did this. The other kind is produced by an experiment in which subjects are instructed to prepare to flex at a prearranged clock time but to refrain from actually flexing and "to veto the developing intention/preparation to act . . . about 100 to 200 ms before [that] time" (Libet, 1985, p. 538).

The results of Libet's veto study suggest an interpretation of type I and type II RPs that is contrary to his own interpretation. To begin to see why, notice that Libet's claim that the subjects in this study veto "*intended* motor action" (1985, p. 38; emphasis added) is implausible (Mele, 1997, p. 322, 2009, pp. 52–53). These subjects were instructed in advance *not* to flex, but to prepare to flex at the prearranged time and to "veto" this. The subjects intentionally complied with the request. They intended from the beginning *not* to flex at the appointed time. So what is indicated by what Libet refers to as "the 'veto' RP" before "about 150–250 ms before the preset time" (Libet, 1985, p. 538)? Presumably, not the acquisition or presence of an *intention* to flex; for then, at some point in time, subjects would have both an intention to flex at the prearranged time and an intention not to flex at that time. And how can a normal agent be in this condition?[3]

A segment of "the 'veto' RP" resembles segments of type I RPs in cases in which subjects do flex, as Libet observes (1985, p. 538). Given that this segment of "the 'veto' RP" is not correlated with a proximal intention to flex, perhaps the similar segments of type I RPs (and of type II RPs) also are not correlated with proximal intentions to flex. Even so, they might be correlated with potential causes of such intentions.

This idea is developed in Mele, 2006 and 2009. The shape the idea takes there is based partly on the following possibilities about subjects in the veto experiment:

> perhaps a subject's wanting to comply with the instructions—including the instruction to prepare to flex at the appointed time—together with his recognition that the time is approaching produces an unconscious urge to flex soon, a pretty reliable causal contributor

[3] Try to imagine that you intend to eat some pie now while also intending not to eat it now. What would you do? Would you reach for it with one hand and grab the reaching hand with your other hand? People who suffer from anarchic hand syndrome sometimes display behavior of this kind. Spence and Frith suggest that these people "have conscious 'intentions to act' [that] are thwarted by . . . 'intentions' to which the patient does not experience conscious access" (1999, p. 24).

> to an urge to flex soon, or the motor preparedness typically associated with such an urge. Things of these kinds are potential causal contributors to the acquisition of proximal intentions to flex. A related possibility is suggested by the observation that "the pattern of brain activity associated with imagining making a movement is very similar to the pattern of activity associated with preparing to make a movement" (Spence & Frith, 1999, p. 27 . . .).[4] The instructions given to [subjects in the veto experiment] would naturally elicit imagining flexing very soon, an event of a kind suitable, in the circumstances, for making a causal contribution to the emergence of a proximal urge to flex (Mele, 2009, p. 55).

The suggestion is that these same items—as opposed to proximal intentions to flex—are candidates for what the pertinent segments of type I RPs signify and that *proximal intentions* to flex emerge later, both in the case of flexes associated with type I RPs and in the case of flexes associated with type II RPs (Mele, 2009, ch. 3). And again, the reaction time study discussed earlier provides independent evidence about when proximal intentions emerge that places their emergence much closer to the muscle burst than −550 ms.

Trevena and Miller conducted a study involving an "always-move" and a "sometimes-move" condition (2010, p. 449). In both the conditions, participants were presented with either an "L" (indicating a left-handed movement) or an "R" (indicating a right-handed movement) and responded to tones emitted at random intervals. In the sometimes-move condition, participants were given the following instructions: "At the start of each trial you will see an L or an R, indicating the hand to be used on that trial. However, you should only make a key press about half the time. Please try not to decide in advance what you will do, but when you hear the tone either tap the key with the required hand as quickly as possible, or make no movement at all" (p. 449). In the always-move condition, participants were always to tap the assigned key as quickly as possible after the tone. Trevena and Miller examined EEG activity for the second preceding the tone and found that mean EEG "amplitudes did not differ among conditions" (p. 450). That is, there were no significant differences among pre-tone EEG amplitudes in the following three conditions: always-move; sometimes-move with movement; sometimes-move without movement. They also found that there was no significant lateralized readiness potential (LRP) before the tone (p. 450). Trevena and Miller plausibly regard these findings as evidence that no part of pre-tone EEG represents a decision to move. The mean time "from the onset of the tone to a key press . . . was 322 ms in the always-move condition and 355 ms in the sometimes-move condition" (p. 450).

Trevena and Miller conducted a second study in which it was up to the subjects which hand to move when they heard the tone. As in the first study, there was an always-move condition and a sometimes-move condition. Trevena and Miller found that, as in the first study, pre-tone EEG "did not discriminate between" trials with movement and trials without movement, "LRP was absent before the tone," and LRP "was significantly positive after the tone for trials in which a movement

[4] Kilner et al. produce evidence that, as they put it, "the readiness potential (RP)—an electrophysiological marker of motor preparation—is present when one is observing someone else's action" (2004, p. 1299).

was made" (p. 453). They conclude, reasonably, that pre-tone EEG "does not necessarily reflect preparation for movement, and that it may instead simply develop as a consequence of some ongoing attention to or involvement with a task requiring occasional spontaneous movements" (p. 454).

Even if Libet's data do not warrant his claim that his subjects have proximal intentions to flex before they think they do, his idea that we have unconscious proximal intentions should not be lightly dismissed. Such intentions may be at work when, for example, experienced drivers flip their turn indicators to signal for turns they are about to make. In a study in which subjects are instructed to flex whenever they feel like it without also being instructed to report after flexing on when they first became aware of an intention, urge, or decision to flex, would they often be conscious of proximal intentions, urges, or decisions to flex? Might unconscious proximal intentions to flex—and, more specifically, proximal intentions of which they are *never* conscious—be at work in producing flexes in the imagined scenario?

Imagine that someone conducts the experiment just sketched and discovers (somehow) that the subjects were never or rarely conscious of proximal urges, intentions, or decisions to flex. Could it legitimately be inferred that, in Libet's own experiment, conscious urges, intentions, and decisions had no effect on the flexing actions? No. One possibility is that some of Libet's subjects treat their initial consciousness of an urge to flex as a go signal. If they do, the conscious urge seemingly has a place in the causal process that issues in the flexing. Another possibility is that some subjects treat the conscious urge as what may be called a *decide signal*—a signal calling for them consciously to decide right then whether to flex right away or to wait a while. If that is so, and if they consciously decide to flex and execute that decision, the conscious urge again seemingly has a place in the causal process, as does the conscious decision. (Notice that the tone in the sometimes-move conditions in Trevena and Miller's studies apparently functions as a decide signal. In the first study, it signals participants to decide whether or not to press the designated key right then; and in the second, it signals them to decide both whether or not to press right then and which key to press, if they decide to press.)

Perhaps it will be suggested that even if a subject treats a conscious urge to flex as a go or decide signal, that urge has no place in the causal process that issues in a flex because an unconscious brain event caused the conscious urge. But the inference here has the same form as the misguided assertion about conscious intention discussed earlier. An x can be among the causes of a y even if the x itself is caused (recall the firecracker example). Possibly, it will be claimed that by the time the conscious urge emerges it is too late for the subject to refrain from acting on it (something that Libet denies) and that is why the conscious urge should not be seen as part of the process at issue, even if subjects think they are treating the urge as a go or decide signal. One way to get evidence about this is to conduct an experiment in which subjects are instructed to flex at a time *t* unless they detect a stop signal—for example, a change in the color of the clock from white to red. (On stop signal experiments, see Logan, 1994.) By varying the interval between the stop signal and the mean time of the completion of a full flex when there is no stop

signal, experimenters can try to ascertain when subjects reach the point of no return. (Time *t* can be a designated point on a Libet clock, and brain activity can be measured backward from *t*.) Perhaps it will be discovered that the point is reached significantly later than time *W*.

13.6 How Accurate Are Subjects' Awareness Reports?

Libet contends that subjects in his main experiment become aware of their proximal intentions well after they acquire them. His primary evidence for the average time of the onset of this awareness comes from the reports subjects make after each flex—reports about where they believe the dot was on the clock when they first became aware of their decision, intention, urge, or whatever, to flex. How accurate are these reports likely to be?

The following labels facilitate discussion:

P-time: The time at which a proximal decision is made or a proximal intention or urge is acquired.

C-time: The time of the onset of the subject's consciousness of an item of the kind just specified.

B-time: The time the subject believes to be *C*-time when responding to the experimenter's question about *C*-time.

Libet contends that average *P*-time is −550 ms for subjects who are regularly encouraged to flex spontaneously and report no "preplanning." And he arrives at an average *C*-time of −150 ms by adding 50 ms to his average *B*-time (−200 ms) to correct for what he believes to be a 50 ms bias in subjects' reports. (For alleged evidence of the existence of this bias, see Libet, 1985, pp. 534–535, 2004, p. 128.) One connection in which *C*-time is important to Libet is his position on veto power. Whether subjects in Libet's studies are ever conscious of relevant proximal urges or intentions early enough to veto them, as he claims, depends partly on what their *C*-times are. The same is true of the question whether, on average, his subjects become conscious of proximal intentions to flex about 400 ms after those intentions emerge in them.

There is an interesting body of work on how accurate *B*-times are likely to be—that is, on how likely it is that they closely approximate *C*-times. This is not surprising. Reading the position of a rapidly revolving dot at a given time is a difficult task, as Wim van de Grind observes (2002, p. 251). The same is true of relating the position of the dot to such an event as the onset of one's consciousness of a proximal intention to flex a wrist. Patrick Haggard notes that "the large number of biases inherent in cross-modal synchronization tasks means that the perceived time of a stimulus may differ dramatically from its actual onset time. There is every reason to believe that purely internal events, such as conscious intentions, are at least as subject to this bias as perceptions of external events" (2006, p. 82).

One fact that has not received sufficient attention in the literature on accuracy is that individuals display great variability of *B*-times across trials. Haggard and Eimer (1999) provide some relevant data. For each of their eight subjects, they locate the median *B*-time and then calculate the mean of the premedian (i.e., "early") *B*-times and the mean of the postmedian (i.e., "late") *B*-times. At the low end of variability by this measure, one subject had mean early and late *B*-times of −231 and −80 ms and another had means of −542 and −351 ms (p. 132). At the high end, one subject's figures were −940 and −4 ms and another's were −984 and −253 ms; and, as I mentioned, these figures are for means, not extremes. These results contribute to grounds for serious doubt that *B*-time closely approximates *C*-time. If there were good reasons to believe that *C*-times vary enormously across trials for the same subject, we might not find enormous variability in a subject's *B*-times worrisome in this connection. But there is good reason to believe this only if there is good reason to believe that *B*-times closely approximate *C*-times; and given the points made about cross-modal synchronization tasks in general and the cross-modal task of subjects in Libet-style experiments, there is not.

Another factor that may make it difficult for subjects to provide *B*-times that closely approximate *C*-times is their uncertainty about exactly what they are experiencing. As Haggard observes, subjects' reports about their intentions "are easily mediated by cognitive strategies, by the subjects' understanding of the experimental situation, and by their folk psychological beliefs about intentions" (2006, p. 81). He also remarks that "the conscious experience of intending is quite thin and evasive" (2005, p. 291). Even if the latter claim is an overstatement and some conscious experiences of intending are robust, the claim may be true of many of the experiences at issue in Libet-style studies. One can well imagine subjects wondering occasionally whether, for example, what they are experiencing is an intention (or urge) to act or merely a thought about when to act or an anticipation of acting soon. Lau and coauthors say that they require their subjects to move a cursor to where they believed the dot on a Libet clock was "when they first felt their *intention* to press the button" (Lau et al., 2007, p. 82; emphasis mine). One should not be surprised if some subjects given such an instruction were occasionally to wonder whether they were experiencing an intention to press or just an *urge* to press, for example. (Presumably, at least some lay folk treat intentions and urges as conceptually distinct, as dictionaries do.) Subjects may also wonder occasionally whether they are actually *feeling* an intention to press or are mistakenly thinking that they feel such an intention.

I argued elsewhere that results reported by Lau et al. (2007) "suggest that reports of *B*-times are reports of estimates that can be influenced by events that follow action" (Mele, 2009, p. 128).[5] A study by Banks and Isham provides confirmation for this claim. They asked subjects in a Libet-style experiment to report, shortly after pressing a response button, where the cursor was on a numbered Libet clock "at the instant they made the decision to respond" (2009, p. 18). "The computer registered the switch closure and emitted a 200-ms beep . . . at 5, 20, 40, or 60 ms after closure." Obviously, subjects were not being asked to report on unconscious decisions; conscious decisions are at issue.

[5] I did not suggest that the estimates are influenced *only* by events that follow action. For evidence that the estimates are also influenced by events that precede action, see Haggard (2011, pp. 19–22).

Banks and Isham found that although the average time between the beep and *B*-time did not differ significantly across beep delays, the following two average times did differ significantly across delays: (1) the time between EMG onset and *B*-time; (2) the time between switch closure and *B*-time. The data display an interesting pattern (see 2009, p. 19):

Beep delay	*B*-time to EMG	*B*-time to beep	*B*-time to switch closure
+5	−21	−127	−122
+20	+4	−124	−104
+40	+4	−135	−95
+60	+21	−137	−77

The beep affected *B*-time, and the beep followed switch closure.

Return to the issue of great variability in *B*-times in the same subject. One way to seek to reduce it is to give the subject a way of conceiving of, for example, making a conscious proximal decision that is easily grasped and applied. Subjects in a Libet-style experiment may be given the following instructions:

> One way to think of deciding to press the button now is as consciously saying "now!" to yourself silently in order to command yourself to press the button at once. Consciously say "now!" silently to yourself whenever you feel like it and then immediately press the button. Look at the clock and try to determine as closely as possible where the dot is when you say "now!" . . . You'll report that location to us after you press the button (Mele, 2009, p. 125).

Subjects can also be regularly reminded to make their decisions "spontaneously"—that is, to make them without thinking in advance about when to press. If, as I predict, subjects given these instructions individually show much less variability in *B*-times than subjects given typical Libet-style instructions, we would have grounds for believing that their reports about when they consciously said "now!" involve *less guesswork* and, accordingly, additional grounds for skepticism about the accuracy of *B*-times in typical studies.

I asked how accurate subjects' reports about when they first became aware of a proximal intention or urge are likely to have been? *Not very* certainly seems to be a safe answer. But there may be ways to improve accuracy.[6] If such *B*-times as have actually been gathered are unreliable indicators of *C*-times, little weight can be put

[6] Would subjects' conscious, silent "now!"s actually express proximal *decisions*? Perhaps not. To see why, consider an imaginary experiment in which subjects are instructed to count—consciously and silently—from 1 to 3 and to press a button just after they consciously say "3" to themselves. Presumably, these instructions would be no less effective at eliciting pressings than the "now!" instructions. In this experiment, the subjects are treating a conscious event—the conscious "3"-saying—as a go signal. (When they say "3," they are not at all uncertain about what to do, and they make no *decision* then to press.) Possibly, in a study in which subjects are given the "now!" instructions, they would not actually make proximal decisions to press but would instead consciously simulate deciding and use the conscious simulation event as a go signal. However, the possibility of simulation is not a special problem for studies featuring the "now!"-saying instructions. In Libet's own studies, some subjects may be treating a conscious experience—for example, their initial consciousness of an urge to flex—as a go signal (see Keller & Heckhausen, 1990, p. 352).

on them in arguments about whether or not there is time enough to veto conscious proximal urges and the like; and the same is true of arguments about whether or not *C*-time is too late for conscious proximal intentions and the like to play a role in producing corresponding overt actions.

13.7 Conclusion

Libet's data do not warrant the claim that his subjects make decisions to move before they are aware of those decisions. Nor do his data warrant the claim that conscious decisions and intentions play no role in generating corresponding overt actions. It is fair to conclude that, on any reasonable conception of free will, the studies and data reviewed here pose no threat to it.[7]

References

Banks, W., & Isham, E. (2009). We infer rather than perceive the moment we decided to act. *Psychological Science, 20*, 17–21.

Bayne, T. (2011). Libet and the case for free will skepticism. In R. Swinburne (Ed.), *Free will and modern science*. Oxford: Oxford University Press.

Haggard, P. (2005). Conscious intention and motor cognition. *Trends in Cognitive Sciences, 9*, 290–95.

Haggard, P. (2006). Conscious intention and the sense of agency. In N. Sebanz & W. Prinz (Eds.), *Disorders of volition*. Cambridge, MA: MIT Press.

Haggard, P. (2011). Does brain science change our view of free will. In R. Swinburne (Ed.), *Free will and modern science*. Oxford: Oxford University Press.

Haggard, P., & Eimer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. *Experimental Brain Research, 126*, 128–33.

Haggard, P., & Magno, E. (1999). Localising awareness of action with transcranial magnetic stimulation. *Experimental Brain Research, 127*, 102–07.

Hallett, M. (2007). Volitional control of movement: The physiology of free will. *Clinical Neurophysiology, 118*, 1179–92.

Keller, I., & Heckhausen, H. (1990). Readiness potentials preceding spontaneous motor acts: Voluntary vs. involuntary control. *Electroencephalography and Clinical Neurophysiology, 76*, 351–61.

Kilner, J., Vargas, C., Duval, S., Blakemore, S., & Sirigu, A. (2004). Motor activation prior to observation of a predicted movement. *Nature Neuroscience, 7*, 1299–1301.

Lau, H., Rogers, R., & Passingham, R. (2007). Manipulating the experienced onset of intention after action execution. *Journal of Cognitive Neuroscience, 19*, 81–90.

[7] Parts of this article derive from Mele, 2009. This article was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed in this publication are my own and do not necessarily reflect the views of the John Templeton Foundation. I am grateful to an audience at a Social Trends Institute Experts Meeting (Barcelona, October, 2010) for comments on a presentation of this article.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *The Behavioral and Brain Sciences, 8*, 529–66.
Libet, B. (1992). The neural time-factor in perception, volition and free will. *Revue de Métaphysique et de Morale, 2*, 255–72.
Libet, B. (1999). Do we have free will? *Journal of Consciousness Studies, 6*, 47–57.
Libet, B. (2001). Consciousness, free action and the brain. *Journal of Consciousness Studies, 8*, 59–65.
Libet, B. (2004). *Mind time*. Cambridge, MA: Harvard University Press.
Libet, B., Wright, E., & Gleason, C. (1982). Readiness potentials preceding unrestricted 'spontaneous' vs. pre-planned voluntary acts. *Electroencephalography and Clinical Neurophysiology, 54*, 322–35.
Logan, G. (1994). On the ability to inhibit thought and action: A users' guide to the stop signal paradigm. In E. Dagenbach & T. Carr (Eds.), *Inhibitory processes in attention, memory, and language*. San Diego: Academic.
Mele, A. (1997). Strength of motivation and being in control: Learning from Libet. *American Philosophical Quarterly, 34*, 319–32.
Mele, A. (2003). *Motivation and agency*. New York: Oxford University Press.
Mele, A. (2006). *Free will and luck*. New York: Oxford University Press.
Mele, A. (2009). *Effective intentions*. New York: Oxford University Press.
Pockett, S. (2006). The neuroscience of movement. In S. Pockett, W. Banks, & S. Gallagher (Eds.), *Does consciousness cause behavior? An investigation of the nature of volition*. Cambridge, MA: MIT Press.
Roediger, H., Goode, M., & Zaromb, F. (2008). Free will and the control of action. In J. Baer, J. Kaufman, & R. Baumeister (Eds.), *Are we free? Psychology and free will*. New York: Oxford University Press.
Spence, S., & Frith, C. (1999). Towards a functional anatomy of volition. *Journal of Consciousness Studies, 6*, 11–29.
Trevena, J., & Miller, J. (2010). Brain preparation before a voluntary action: Evidence against unconscious movement initiation. *Consciousness and Cognition, 19*, 447–56.
van de Grind, W. (2002). Physical, neural, and mental timing. *Consciousness and Cognition, 11*, 241–64.
Wegner, D. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press.

Chapter 14
Towards Non-physical Realism

Jean Staune

Abstract The objective of this paper is to show that a dualist model is not only possible but also most appropriate in order to understand the problem of consciousness and the existence of Free Will. By "dualist" we refer to the assumption that reality cannot be explained exclusively by observable causes in space-time. The dualist view we speak about here does not totally correspond to the classical conception of "dualism," according to which matter and consciousness would be two radically separate things.

The first part of this paper is dedicated to EPR-type experiments which show that no matter what the interpretations are, we are obliged to call into question the classical notions of time and space and obliged to accept that ultimate reality cannot be localized in or be dependent on time and space.

In the second part I will be arguing that in order to be understood the experiments of Benjamin Libet must be studied in a dualist framework, even though Libet was not himself a dualist. A Copernican revolution is therefore possible not only in our understanding of the world but also in our comprehension of the nature of consciousness.

Keywords Mind-body problem • Free Will • Non locality • Dualism • Libet • EPR • d'Espagnat

14.1 Introduction

I will not repeat here a description of what Non-Locality is, nor of the experiments which have permitted its existence to be proven, thereby giving reason to Niels Bohr in the debate that opposed him to Albert Einstein. Indeed, it has been

J. Staune (✉)
Université Interdisciplinaire de Paris, Paris, France
e-mail: staune@uip.edu, http://www.staune.fr/

A. Suarez and P. Adams (eds.), *Is Science Compatible with Free Will?: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience*,
DOI 10.1007/978-1-4614-5212-6_14,

presented in details in this book by Nicolas Gisin. I will attempt instead to show the philosophical implications that we can draw from these experimental results, results which belong to the most important ones of the twentieth century.

There are, however, two ways of approaching the phenomenon. Either, as Bell puts it, there is an influence which is not subject to matter or energy (because otherwise it could not travel faster than the speed of light) and which goes from one particle to the other, in which case we speak about "Non-Locality" as it violates the Locality Principle as conceived by Einstein, or, as the majority of today's physicists think, the two particles *form one and the same object* even if they are measured by instruments in theory thousands of miles apart. In this case we would be talking more of "Non-Separability" because the two particles cannot be separated (when they have not been measured). After the "before-before" experiment realized by the team of Gisin, this second interpretation seems the most probable.

As Bernard d'Espagnat says "as far as Non-Separability is concerned the two descriptions are equivalent. In either cases, a violation of Einsteinian separability necessitates an instantaneous interaction at a distance, either between two distinct systems or within a single and same system spread out over space" (d'Espagnat 1980, p. 86).

We can see in both cases that there is no possible escape route: we are led to radically revise our beliefs about the very foundations of reality.

This is why this result is of such importance: it represents a major shift in our knowledge. This experiment jettisons a great number of different views of the world, which can be thrown out like used bus tickets.

14.2 What Does Non-locality Imply About the Nature of Reality?

The standard interpretation of Quantum Mechanics is named the "Copenhagen interpretation" attributed to Niels Bohr's influence. The majority of today's physicists have adopted this view because it allows them to avoid asking any philosophical questions. Bohr tells us, "Quantum Mechanics deals not with reality but with what we know about it" (quoted by Ortoli and Pharabod 1984, p. 83). "Quantum Mechanics simply allows those observers with measuring apparatus to predict their observations correctly. There is no point in seeking to explain *why* it works. It is enough to see that it works and to apply its formalism" (Ortoli and Pharabod 1984, p. 83).

In other words Quantum Physics predicts experimental results, but there is no point in trying to represent the reality which might exist (or which might not exist) behind the phenomena observed. This certainly avoids a lot of headache but for those who wish to understand the nature of the world, the Copenhagen interpretation is, in Etienne Klein's words, "designed to frustrate people."

Some devotees of the interpretation such as Pascual Jordan are quite idealistic. He actually goes as far as removing any meaning to the question of the existence of any sort of reality. "A common error, from a positivist view, is to deny the existence of the exterior world. The negation of a proposition devoid of meaning is a proposition devoid of sense. The idea of the non-existence of a real exterior world has no more meaning than its existence. Neither one nor the other is true or false, they are completely meaningless" (Jordan 1936, p. 309).

It is not necessary, however, to adopt such an extreme position to have difficulties with the notion of reality. The Copenhagen interpretation does not allow us to speak about the existence of the electron (let alone of its properties) when it is not under observation, which is close to idealism. As Bernard d'Espagnat says, there is a certain ambiguity to the position of many physicists who claim to support it: "The majority of physicists are happy to use Quantum Mechanics without bothering to question its basic rules. How can they justify this? According to them, these fundamentals have long since been elucidated by the Copenhagen school. Even those physicists who regard themselves as realists are quite prepared to take this stance. Do they actually realize to what extent they are distancing themselves from all realism—or materialism—in the accepted sense of these terms? Heisenberg tends to agree with Kant. This means that the realism of those physicists, who rely on their elders' views without questioning the fundamentals, is akin to that of the philosophy known as Kantian idealism. Are my esteemed colleagues fully aware of the slant this gives to their ideas, and if they are, are they prepared to admit this to their students or to their public?" (d'Espagnat 1982, p. 59).

This stance (which could be crudely called "shut-up and calculate!") was illustrated by Richard Feynman, Nobel Prize recipient for Physics, in his allegory concerning the Mayas. The Mayas knew how to predict eclipses but did not have the slightest idea of the real nature of the sun or the moon. Let's suppose, Feynman says, that a Mayan student tells his tutor to imagine that the earth, the moon and the sun are three balls floating in space thereby explaining the eclipses.

> "Can you predict anything else other than what has already been predicted?" the tutor asks the student.
>
> No.
>
> Then your theory is useless because only experiments count. Being interested in the nature of things goes beyond the realms of science and borders on metaphysics! (Feynman 1979, p. 202)

14.3 Non-physical Realism

So what is left for those who want to go beyond idealism? There is of course a realist type of stand, but it corresponds to a "non-physical realism," the antithesis of classical realist thinking which has become associated with materialism.

Let's take the example of a rainbow (d'Espagnat 2002, pp. 398–402). You might think on seeing it for the first time that it is a solid object whose two extremities touch the ground. Then you notice that when you move, it moves with you. Does this mean

that the rainbow is an illusion and a figment of your imagination? No, of course not. Its existence is dependent on the presence of water droplets in the atmosphere and the refraction of the light's rays. Nonetheless, certain important characteristics of the rainbow, such as its position and speed, are dependent on you and where you stand. The situation is identical for all elementary particles and even for atoms, in the conception of non-physical realism. These are not figments of our imagination, but some of their essential characteristics depend on the way we observe them. This introduces a radical difference with science's normal goal, as summarized here by Albert Messiah, "There is a fundamental premise at the outset of any scientific enterprise, that nature possesses an objective reality, independent of our sensory perceptions or our investigative means. The point of physical theory is to give an intelligible account of this objective reality" (quoted by d'Espagnat 1979, p. 59).

For example, if we are told that "gravitation only depends on mass and the square of distance," we are talking about "a very objective statement," because mass and the positions of macroscopic objects do not vary when measured. In this field, this type of statement is referred to as having "strong objectivity." The statements made in Quantum theory, however, refer to our perceptions or to our instruments. They are objective only in as much as they hold true for any observer. *We cannot therefore say that they are absolutely true* because their truth needs to relate to the community of human observers. They are statements of "weak objectivity" (d'Espagnat 1979, p. 60).

Quantum Mechanics cannot therefore describe what is real in terms of strong objectivity. This is why even if physical realism or classical realism abandons the materialist claim of describing the foundation of what is, as being constituted of objects, it cannot be compatible with this type of physics.

Another characteristic of this new form of realism is its "distant" nature. Not distant in a geographic sense, but "conceptually distant" because concepts that we are familiar with, those which are close to our way of understanding things, are no longer applicable. We could refer to a "strange realism."

To see to what extent it is strange, let us analyze the de Broglie paradox. One electron is placed in a vacuum box (this is a "thought" experiment, but technology today enables us to keep an electron inside a magnetic field without it interfering with other bodies). The box is cut into two, one half is sent to Tokyo and the other to Paris. On opening the Paris box, the electron is revealed. Franco Selleri, one of the few physicists who does not accept the case today for Quantum Mechanics tells us, "if we open the Paris box and find the electron in it, the natural reaction of most physicists will be to say that the electron observed in Paris at the time of opening the box was also there before the box was opened" (Selleri 1986) and that therefore the half-box in Tokyo was empty right from the beginning. Perhaps that is the "natural reaction," but it is not the right reaction if the matter is given some thought! If we follow Quantum Physics, the electron might be spread throughout the box. When the box is cut into two (putting aside the fact that this action will have perturbed the wave function), the electron will have spread into both halves of the boxes. When the electron is observed in Paris, there is a reduction in the wave packet and the probability of observing the electron in Tokyo is only eliminated at this point.

Thus with Quantum Mechanics—contrary to common sense—we can say that the fact that the electron has been observed in Paris does in no way imply that it was already present in the half-box in Paris before observation and therefore that the box in Tokyo was empty. Before opening the boxes, the electron was in a state of superposition, as in: "the electron is in Paris" and "the electron is in Tokyo."

In this new conception of a "strange" and "distant" realism, compared to the principles which rule our daily lives, this paradox should not surprise us! However I can perfectly well see how troubling the following statement is: "when the electron is found in the box we cannot claim that it was already in Paris before the box was opened."

As Albert Messiah has said, realism supposes the existence of a reality independent of our perceptions and our means of observation. What Quantum Physics does, is to show that if such an independent reality exists, it is *not* the physical reality that we can see, touch, feel or measure! Indeed, this reality—like the rainbow—is not independent of our perceptions and our means of observation. Nonetheless the experiments that we have just described do show us that something escapes not only time and space but equally matter and even energy. This "something" is a good candidate for independent realism but should, however, be considered as *non-physical or distant*. This independent reality cannot be described by science. It can at best be very imperfectly approached by a science of *weak objectivity*—but not strong objectivity.

This conception of realism has been studied in depth by d'Espagnat (1979, 1994, 2002). Different views which also lead to the rejection of all classical materialist conception of reality and which underline the non-ontological nature of the world we live in, have also been expressed by several other physicists: Raymond Chiao, Olivier Costa de Beauregard, Paul Davies, Amit Goswami, Andreï Grib, Menas Kafatos, Stanley Klein, Thierry Magnin, Alexis Nesteruk, Basarab Nicolescu, Lothar Schäfer and Henry Stapp amongst others.

To best understand how this new conception of reality differs from the old one, it is worth thinking about Bernard d'Espagnat's message, "one of the teachings of modern science of so called 'matter' is the following: the 'thing', if there is one, which remains preserved is not concrete but abstract. It is not something which is close to the senses but which, on the contrary, is a pure mathematically abstract number such as theoretical physics has revealed to us. In other terms, compared to our senses and the concepts that are familiar to us, reality is undeniably distant. In order to do justice to this very important discovery, when we speak about it, I think that it is crucial to know that the word 'matter' is the wrong one and that the more appropriate word 'Being' should be reintroduced" (d'Espagnat 1982, p. 55).

What an extraordinary conceptual change, when what is considered as real is in fact abstract, not concrete and closer to mathematical formulae than to a grain of sand! Quite the opposite of all the scientific and materialist conceptions of the previous centuries!

Can we escape the conclusion that ultimate reality (if it exists, as it is of course impossible to refute idealism) is *not* imbedded in space, time, matter and energy? I think that it is impossible; actually, all the different models that attempt to

establish a strong realism are non-local, as shown by the principal among them, developed by Louis de Broglie, David Bohm and John Bell (sometimes called the BBB model). Indeed, the Quantum potential existing in this model is, by definition, non-local. So it is a model where the very structure of things escapes time and space. The model of parallel universes, invoked by materialists in numerous circumstances to dispose of embarrassing concepts such as the Anthropic Principle, is, here, of no help. This model certainly doesn't enable to reduce Non-Locality to a phenomenon that would happen in time and space, regardless of which given universe. Thereby these models—also not credible at the experimental level—if they can, in theory, restore strong objectivity and make the notion of weak objectivity disappear, cannot, by nature, be such as to restore classical realism; any new theory will be non-local, such as very well concluded by Nicolas Gisin: "physics offers no story in space and time to explain or describe how these correlations happen. Hence, somehow, non-local correlations emerge from outside space-time" (Gisin 2012, Chap 3).

It is important to note that this point is supported by one of the principal representatives of the most extreme materialistic trend in the field of Physics, Jean Bricmont: he compares indeed Non-Locality to a magician's act (he does at least stipulate that this is just an analogy!) capable of acting from a distance on a person by manipulating his effigy whatever the distance separating them. He says rightly that information cannot be transmitted using Non-Locality, "but other disconcerting aspects are there, such as instantaneousness, individuality, effects which do not decrease with distance."

Bricmont wastes no time, however, in qualifying this action ("aspects which do not decrease with the distance, contrary to all known physical forces, which propagate more rapidly than the speed of light") by speaking of the " 'magical' properties of Non Locality" (Bricmont 1995, pp. 150–151).

And he concludes: "Non Locality is a property of nature established by means of experiments and elementary reasoning, independent of the interpretation of formalized Quantum Physics. As a result any other theory which might replace Quantum Mechanics will also be non-local" (Bricmont 1995, pp. 131–179).

But numerous professional meetings that I've had over the last 20 years with physicists, as well as the conferences I've organized in the field of Quantum Physics, have shown me that many physicists were simply not aware of the conceptual leap that was involved in these experimental results. In some extreme cases, some were aware of it, but refused to believe in it (in the mid-1990s, I even met a UCLA professor of Physics who told me: "if these results are confirmed, I will abandon my position and do pottery"). As Thomas Kuhn showed, the changes of paradigms never happen easily and it is often most difficult for a community to accept a new paradigm taking place in their field.

When modern ideas started to emerge, the Inquisition tried to stop them from spreading. The resulting religious obscurantism has been widely chronicled. Today we are living through the same situation. This time, those who find themselves in a dominant position and who see their position weakened in the face of scientific progress are the materialists and those in favor of scientism. Today, in most

Western countries, obscurantism no longer has religious connotations (the Petit Robert dictionary definition for Obscurantism is: preventing the spread of knowledge or culture amongst a population), but materialist connotations (it does not concern the materialists as a whole, far from it, the same way that religious obscurantism at its height did not affect all clerics).

Obscurantism comes in three guises:

- Omission: writing a book on the nature of reality as a physicist and not talking about the EPR paradox.
- Reassurance for the wrong reasons: talking about these issues by saying that indeed peculiar things happened but that everything has settled down and that "common sense," i.e., our classical concepts, are no longer threatened.
- Misinformation, pure and simple: to say something that is inexact about a subject that the author is expected to know about.

A particularly interesting example in this field is the one of Nobel Prize in Physics, Murray Gell-Mann in his famous book *The Quark and the Jaguar*. In a chapter dedicated to the Einstein–Podolsky–Rosen paradox and to the experiences stemming from it, he doesn't hesitate to write that "the principal distortion (concerning these experiences) spread in the medias and in various books, is the affirmation, implicitly or explicitly accepted, that to measure the polarization of one of the photons affects, in one way or the other, the second photon. In fact, the measurement causes no propagation of any physical effect, from one photon to the other. . . On each branch, the situation resembles Bertlemann's socks described by John Bell in one of his papers. Bertlemann is a mathematician who always wears a sock of pink color and the other of green color. If you only see one of his feet with a green sock, you will know that his other foot wears a pink sock. And yet no signal has been propagated from one foot to the other. In the same way, in the experiment confirming Quantum Mechanics, no signal passes from one photon to the other; there is no action taking place from a distance. This wrong allegation, according to which the measurement of one photon immediately affects the other, leads to all kind of unfortunate conclusions" (Gell-Mann 1985, pp. 196–197). What is especially unfortunate is to see a Nobel Prize Winner in Physics say something so seriously inaccurate. In fact this is inaccurate in two respects: first, it is clear that the measurement made on one element of the system affects the entire system, it is the very essence of Non-Locality; then, Gell-Mann makes reference to a famous chapter by John Bell as a support to his thesis, while Bell states, in this very chapter, in three different places, exactly the contrary of what Murray Gell-Mann implies: "We cannot avoid the fact that intervention on one side has a causal effect on the other" (Bell 1987, p. 150), "Some correlations are locally inexplicable. They cannot be explained without action at a distance" (Bell 1987, p. 152), "For the experiment described, that would not only be a mysterious long distance influence (a non-locality, or an action at a distance in the weakest sense) but an influence propagating more rapidly than the speed of light, a non-locality in the strictest and most difficult to accept form" (Bell 1987, p. 153).

We are here in a quite extraordinary situation at the epistemological level. Let's analyze it:

- Murray Gell-Mann has by far the scientific level to understand John Bell's paper "Bertlemann's Socks and the Nature of Reality," as well as to understand what experiences such as Aspect's prove or disprove.
- Murray Gell-Mann is of good faith; I support here the postulate according to which he is not looking to deliberately misguide his readers.
- Murray Gell-Mann makes the analogy developed by John Bell say exactly the contrary of what it says in three different passages; even worse, Gell-Mann claims that the experiments prove the model that they precisely refute; indeed, what the experiments refute is that *before* the entrance of Mr. Bertlemann in the room, meaning before the measurement, the color of the socks is already determined, pink for one and green for the other (meaning that the polarization of the photons is determined *before* the measurement, the very point which is refuted by Alain Aspect's experiments!). The experiments oblige us to accept a model where the color of the socks would be randomly fixed when we see them (meaning when they enter in the room) and where and when the first sock is green, we know in advance that the other will be pink. I have decided to call this quite extraordinary phenomenon the "Gell-Mann effect." As we will see, the Gell-Mann effect can play an ever more important role in the field of neurosciences and the question of Free Will. For now, we must simply retain that it is difficult for a new paradigm to make its way through, even among specialists of this question.

To understand the extent to which such phenomena are widespread among physicists, let's get back to Jean Bricmont, whose extremely materialistic sentiments don't impede him to be clear sighted in regards to the extraordinary implications of the experiments which brought to light Non-Locality. It so happens that Jean Bricmont is the co-author, with Alain Sokal, also Quantum physicist, of a book that follows a hoax on the part of Alain Sokal, aiming to denounce the absurdities of some philosophers and sociologist supporters of relativism, who succeeded in publishing a completely absurd article in a referee journal.

In this parody he states that "an observation made here and now can not only affect the observed object but can also affect another object as far away as one wishes from the first. This phenomenon, which Einstein called 'phantomatic', incurs a radical re-evaluation of traditional mechanist concepts of space, object, causality and suggests an alternative view in which the universe is characterized by interconnection and holism" (quoted by Sokal and Bricmont 1997, p. 217).

Realizing this was a parody, the reader could well believe it to be exaggerated or false, but not only is it rigorously exact but the words are rather less daring than those of Jean Bricmont when he writes about the same phenomenon in all seriousness!

As a conclusion to this first part, we can state that:

- Either the present tendency concerning the impossibility to establish a Quantum theory of strong objectivity is confirmed, and in this case we must admit the existence of another level of reality.

– Either we'll be able one day to establish a theory of a Quantum Potential type, but even in this case, reality will have to integrate a dimension that is beyond time and space, as well shown by David Bohm's development of reflections on his own model (he compared both particles in a situation similar to EPR to two fishes situated on two television screens, whose movements were perfectly coordinated since it was in fact two images of the same fish, filmed by two cameras—model which could not introduce more clearly another dimension in our reality).
– All this constitutes, as we have seen, a true new Copernican revolution, susceptible to change many dogmas in the Mind-Body problem. The immense majority of neurologists consider that the brain produces consciousness, meaning that it contains in itself all necessary elements in order to fabricate it. But let's imagine that the brain is not an IPod that contains the music it can play, but a radio retransmitting what it receives.

14.4 Is the Brain an IPod or a Radio?

Imagine that extraterrestrials have been observing our behavior for years and not wishing to disturb us, they make sure that we are not aware of them. They take advantage of our holidays to enter our homes and study objects which they find there. Just imagine them in a teenager's bedroom. Looking at the CDs and the stereo equipment makes them soon realize that sounds are encoded in digital form which the player decodes to reconstitute the sound. Looking at an IPod will lead to the same conclusion. If the storage system is technologically more advanced and can stock a greater quantity of sounds, the techniques for storage and reading won't be any more difficult to understand because of this. Looking at the radio will, however, cause great confusion. Where are the sounds stocked in the radio and how are they read? To try to understand it, they will perform several experiments by altering or removing parts of the radio. They will notice that the sound emitted is either modified or non-existent. They will logically conclude that even if the radio is a very technically advanced object, its overall principle is not very different from that of an IPod or the CD player in that it emits sounds which are stocked within it. They will be so convinced of this conclusion that if one day they take these objects back to space and see that the radio does not work, whereas the IPod carries on working, they will probably assume that the radio is more sophisticated and therefore more sensitive to the magnetic field of their spaceship or to the effects of gravity.

Finally they will castigate any theory suggesting that the sounds are not stocked in the radio but are emitted by some mysterious source and treat such suggestions as "prehistoric," "magical" or "mystical."

The point is that today there is no proof that the brain is the equivalent of an IPod or a CD and nothing prevents it from being a radio.

Consciousness is modified when certain areas of the brain are modified but this does no more prove that the brain *produces* consciousness than the fact that music is different when the components of the radio are modified proves that the radio *produces* the music. A minority of neurologists have no hesitation in going

further and considering the brain to be a *condition* and not the ultimate cause of consciousness. But they are confronted with the famous question: how can the mind influence the brain without violating physical laws, the first of which would be the law of energy conservation?

Thanks to Frederick Beck, a Quantum physicist and director of the Department of Theoretical Physics of the University of Darmstadt, the famous neurologist John Eccles found the solution.

Eccles received the Nobel Prize for his work on the functioning of the synapse. The synapse is the essential element for the transfer of the nerve impulse from one neuron to another. This transfer depends on exocytosis, i.e., the bursting of a small number of vesicles each containing 5–10,000 molecules of neuron transmitters. The opening of each vesicle works in an either "all or nothing" way and depends on the displacement of a miniscule part of the membrane of the vesicle (weighing 10^{-18} g). When the nerve impulse arrives in the synaptic button at the end of the axon, the exocytosis allowing the transmission of the "message" to the following neuron has usually only a one in three or four possibility of happening.

By doing a quantum treatment of exocytosis, Beck has shown that the probability of this event happening could be increased or diminished without it constituting a violation of the law of energy conservation, because the masses involved in the phenomenon of exocytosis are small enough to be part of the uncertainties existing on a quantum level.

Beck's and Eccles' work was published by the American Academy of Science (Beck and Eccles 1992) and is a very important piece of work. It does not prove that the mind acts on the brain; it shows that it is *theoretically* possible. Since 1992, therefore the main obstacle to the acceptance of a dualist view no longer exists and has resurfaced as a possibility on a scientific level.

Thereby Quantum Mechanics allows at a scientific level the conception of a dualist vision of the relationship between the body and the mind: first by showing that there exists another level of reality or another dimension, susceptible to house a non-material entity such as the mind. Secondly, by showing that it is theoretically possible that a non-material entity can influence the behavior of a material entity such as the brain. Suarez (2013) proposes in this book a very interesting hypothesis which, it will be fascinating to see if, as he pretends, can be tested at the biological level, at least in an indirect manner. But are there other evidences in favor of such a dualist conception?

Contrary to Mele's view (2013), I think that the experiments on the Readiness Potential give us an interesting lead, while being relevant to the heart of this book's subject matter, Free Will. Although Mele mentions this experiment in his paper, please allow me here to review it briefly.

Hans Helmut Kornhuber and Lüder Deecke discovered in 1964–1965 that about 1 s before a subject has made a gesture, a potential called the Readiness Potential appears in the supplementary motor area. However you do not have the impression that a second has passed between the moment you decide to press the button and the moment you perform the gesture. Libet sheds light on the situation with the following experiment (Libet 2004, pp. 123–156):

The subject is seated in front of a disc on which a black point rotates at a speed of two rotations per seconds. The subject can decide to press on the button from time to time as he wishes. He must say "When I decided to press on the button the black spot was on X." During this time the potentials which are produced in the supplementary motor area are registered. It is noted that the readiness potential begins 0.55 seconds before the act is effected but that the subject reports that he decided to press on the button at the moment of the readiness potential was at its maximum, that is, 0.2 seconds before the act of pressing the button. The act takes place and a discharge of potential takes place, signaling that the gesture has been performed. This is an important detail. Moreover, the readiness potential develops initially in the two hemispheres despite the fact that in the end only one hand moves. It "lateralises" around 0.2 seconds before the act, that is, it disappears from the hemisphere corresponding to the hand that will not move but develops in the other hemisphere.

Materialists were thrilled at the result, stating: "this is the proof that Free Will does not exist. When we think that we have made the decision to press on the button, our brain has in fact already decided 0.35 s beforehand without us even being aware of it!"

Libet did not stop here. He identified subjects who in the end did not move the hand. When the subject is asked what happened at this moment, he says that he was about to hit the button but changed his mind. This led Libet to perform other experiments where the subjects were instructed to act at a prearranged time but to veto this. He concluded: "Subjects can in fact 'veto' motor performance during a 100–200 ms period before a prearranged time to act" (Libet 1985).

Something fundamental happens 0.2 s before the act. This is the moment where the "I" or the "self" has a chance to *stop or to continue* the processes which have been started without it.

This corresponds to our everyday experience. We make a lot of movements without really being aware of them. It is the case of hand movements during lively discussions. We can, however, "take control" at any moment of our bodies by crossing our arms and keeping our hands still.

Free Will is no illusion then. But it is more limited than we thought. It can veto potential acts which we have not initiated ourselves.

An apposite metaphor is that of the football referee. A whole match can be filmed by filming the ball in close up. What is a football match?

A reductionist like Changeux might say, "It is nothing more than twenty two pairs of feet and four hands hitting a ball."

But there is an extra ingredient, namely the referee.

– How come? I have watched dozens of football matches (in close-up) and I have never seen a referee? Does your referee kick the ball too?
– No but . . ...
– Then do not go telling outrageously non-scientific stories, your referee has no role to play in a football match, in fact he probably doesn't exist at all.

At the end of the match, however, it is usually the referee and not the players who is hit with cans by the supporters, proof indeed that he plays an essential part in the match. His role is to let the players play except for the rare moments when he whistles.

Replace "referee" by "mind" and reread the text again and you will understand why this experiment of Libet's is crucial. We cannot "objectivise that which is not an object." We cannot see the mind but we can indirectly deduce the existence of something which affects the neuronal process because some readiness potential is aborted, just as one can deduce the existence of a referee by observing that the players all stop at the same time at certain moments during the match.

As shown by Alfred Mele's paper, this interpretation of Libet has been questioned; nonetheless, and until proven contrary, I think that the interpretation given by Libet is the correct one and that it constitutes the most fascinating scientific support in favor of the existence of Free Will. First, the lateralization of the readiness potential shows us that something crucial happens 0.2 s before the action. Second, the model of Libet corresponds well to our intuitive apprehension: we are sometimes capable of doing elaborated gestures, such as driving a car, without being conscious about it. And we have all experimented this famous veto, a day for example where we were about to do a gesture but kept our arm from moving at the last second.

Does this veto phenomenon correspond to the realization of another readiness potential that would be symmetrical to the first one (which means that it will reach its term when there is veto, and which would abort as the movement reaches its term)? Libet mentions that: "There is no experimental evidence against the possibility that the control process may appear without specific development by prior unconscious processes" (Libet 2004, pp. 45–46). If this not prove in any ways the existence of Free Will it gives support to it as Libet said: "My conclusion about Free Will, one genuinely free in the non determined sense, is then that its existence is at least as good, if not a better scientific option than is its denial by determinist theory" (Libet 1999).

The movements done by the subjects are movements that are totally insignificant. Let's imagine now to couple this experiment with Milgram's experiment and test the moment when the subject will push a button thinking that he could provoke the death of a tested subject.[1] Can we not think that in such a case the EEG would be completely different?

In all cases, I think fundamental to retain that Libet's experiment causes less problems if we are in a dualist framework; it maybe the reason why this sort of work has been criticized, instead of being explored, despite the fascinating opportunities offered for a better knowledge of Free Will.

14.5 The Great Scientific Return of a Dualist Conception of the Mind-Body Problem

The dualist explanation has a bad reputation.

It is "fundamentally anti-scientific," "It must be avoided at all cost," and "accepting dualism is renunciation" (Dennett 1993, pp. 54–55).

[1] This idea is the result of a conversation about this subject with Jean-François Lambert.

Practically every neuroscience book starts with a criticism or two about dualist conceptions, but why not have a closer look?

- We have seen that phenomena such as Non-Separability can have a causal effect on our world and yet is neither composed of matter nor energy.
- At the very least, our journey across Quantum Physics leads us to the conclusion that what exists is not limited to those things that are included in time and space nor comprise matter and energy.
- Doesn't this constitute an argument for the possible existence of a non-localized mind neither comprises matter nor energy?
- Since the publication of the Beck and Eccles article in 1992 (which to my knowledge has never been criticized by any publication in a referee journal) the main theoretical obstacle to a dualist conception of the mind has disappeared.
- Even Descartes could not have dreamt that science could one day provide a framework for such beliefs.
- Isn't the dualist model the most logical solution to Libet's other extraordinary experiments, demonstrating that consciousness can go backwards in time and therefore is not totally situated in time? Libet is not a dualist but he does take care to mention that nothing forbids the existence of a Cartesian type of dualism (Libet 2004, p. 221).
- Remember that many famous scientists share the view that the brain and the mind are two identical things, a view that is undeniably refuted by Libet's experiments whilst they still explain that dualism is anti-scientific. This is a nice illustration of the parable of the straw and the beam.
- Isn't the dualist model the best explanation of the fact that split-brain subjects retain a unique identity?
- Isn't the dualist model the best explanation when one sees that an instance can, at the crucial moment, stop the processes that have been initiated unconsciously in the brain, thus supporting the existence of Free Will?
- Isn't the dualist model the best explanation when one sees that the *intention* to do something can have some physical consequences on the brain and even on the immune system?
- Isn't the dualist model also as good an explanation as that of theory of emergence that the mental states can be radically different from the associated neuronal states?

It therefore seems difficult to reject the hypothesis that the dualist model would provide the best theoretical framework for developing future research on the nature of human consciousness, when purely scientific facts are taken into consideration.

When non-material entities are mentioned, such as the mind or archetypes, etc., materialists immediately retort that this is a way of ossifying the research since, instead of researching a physical cause, something that cannot be verified is being postulated.

But here, the exact opposite is true!

The following are areas of research most likely to bring real progress in the realm of consciousness science:

- The development of Libet's approach, involving the possibility that consciousness can extract itself from time (even if just for a little while).
  Empirical confirmation of this issue lies perhaps in examples such as road accidents where some witnesses have reported that a moment that has only lasted 3 s ("I saw the lorry and crashed into it") seemed to last for 30. It is as if consciousness escaped the bounds of time in order to have more reaction time. This is just one more lead for research amongst many others.
- Research on several cases where we have a hint of the existence of an "operator inside the brain which does not limit itself to the sum of its parts" as Jean François Lambert puts it and which can either stop the processes initiated unconsciously by the brain or stimulate physical processes in the brain uniquely by thought.
- Research on subjects which are currently taboo such as Near Death Experience which shows us that incredible discoveries about human nature are still to come.[2]

I see no foreseeable research that is as promising as that just described within the current paradigm of neurosciences for which "consciousness is a product of the brain," yet by means which we have no ideas.

Today the dualist model is the richest hypothesis for explaining facts that are the products of research in neurosciences. In this field however (just as in that of Evolutionary Science where thousands of researchers study the fruit-fly which has not really evolved over 50 million years, in the hope of understanding the mechanisms of Evolution), a paradigm forbids the openness to all non-physical reality, thus blocking the potentially most fruitful research, although this taboo no longer exists in the realms of Physics, Astrophysics and Mathematics where it is demonstrated that several layers of reality coexist, beyond the boundaries of time, space, energy and matter.

It is important to note that the dualist view we speak about here does not totally correspond to the classical conception of "dualism," according to which matter and consciousness would be two radically separate things.

What we have seen in the part devoted to Physics incites us to think that the belief most in harmony with our knowledge is that consciousness and matter stem from a unique substance which would "ante-date the scission between the subject and the object," according to Bernard d'Espagnat and to be found beyond space, time, and energy. In other words, consciousness and Free Will refer to processes which are not completely in space time.

[2] Some research shows that the "out of body experiences" described by some people who have been close to death might not be an illusion as is generally thought (Van Lommel et al. 2001; Sabom and Kreutzinger 1978; Sabom 1983). Benjamin Libet himself has refined a protocol for testing the reality of this phenomenon in a rigorous fashion (Libet 2004, p. 216–219).

14.6 Conclusion

A convergence seems to draw itself out between, on one hand, concepts provided by Quantum Mechanics, and experiences made in Neurosciences on the other. Quantum Mechanics shows us that ultimate reality isn't limited to the familiar dimensions of time and space, energy and matter. Experiments such as Benjamin Libet's not only on Free Will, but also on the Backward Referral in Time, show that a dualist conception (theoretically allowed by models such as Beck's and Eccles') is more probable than a monist conception. Among these convergences, we must note the existing one between the model drawn from Quantum Physics by Antoine Suarez in this book and the one proposed by Benjamin Libet from his experiment on Free Will. Suarez's model would explain why it is *necessary* that a large amount of our everyday gestures are unconscious; it is the price to pay in order for conscious and voluntary gestures to be possible. I think that such a model would certainly have interested Benjamin Libet. Of course, the general coherence that comes out of such convergences does in no way constitute a proof for the validity of these approaches; nonetheless it should invite numerous researchers to work in the field of the possible interaction between Quantum Mechanics and the Mind-Body problem, following those who have opened this path such as Roger Penrose, Sir John Eccles, Henry Stapp, Mario Beauregard, and today Antoine Suarez. But we have to be aware that in order to do so, we will have to overcome a very strong and powerful "Gell-Mann effect", this time in the field of Neurosciences.

References

Beck, F, & Eccles, J. (1992). Quantum aspects of brain activity and the role of consciousness. *Proceedings of the National Academy of Sciences of the USA*, *89*, 11357–11361.

Bell, J. (1987). *Speakable and unspeakable in quantum mechanics*. Cambridge: Cambridge University Press.

Bricmont, J. (1995). Contre la philosophie de la mécanique quantique. In R. Frank (Ed.), *Les sciences et la philosophie. Quatorze essais de rapprochement*. Paris: Vrin.

Dennett, D. (1993). *La conscience expliquée*. Paris: Odile Jacob.

d'Espagnat, B. (1979). *A la recherche du réel*. Paris: Gauthier-Villars.

d'Espagnat, B. (1980). Théorie quantique et réalité. *Pour la Science*, *27*, 72–87

d'Espagnat, B. (1982). *Un atome de sagesse*. Paris: Seuil.

d'Espagnat B. (1994). *Le réel voilé*. Paris: Fayard.

d'Espagnat, B. (2002). *Traité de physique et de philosophie*. Paris: Fayard.

Feynman, R. (1979). *La nature de la physique*. Paris: Seuil.

Gell-Mann, M. (1985). *Le Quark et le Jaguar*. Paris: Albin Michel.

Gisin, N. (2013). Are there quantum effects coming from outside space-time? Nonlocality, freewill and "no many worlds". In A. Suarez & P. Adams (Eds.), *Is science compatible with free will?* New York: Springer. Chapter 3.

Jordan, P. (1936). *Anschauliche Quantentheorie*. Berlin: Springer.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *The Behavioral and Brain Sciences, 8*, 529–566.

Libet, B. (1999). Do we have free will? *Journal of Consciousness Studies, 6*(8–9), 47–57.

Libet, B. (2004). *Mind time*. Cambridge: Harvard University Press.
Mele, A. (2013). Free will and neuroscience: Revisiting Libet's studies. In A. Suarez & P. Adams (Eds.), *Is science compatible with free will?* New York: Springer. Chapter 13.
Ortoli, S., & Pharabod, J.-P. (1984). *Le cantique des quantiques*. Paris: La Découverte.
Sabom, M. (1983). *Souvenirs de la mort*. Paris: Robert Laffont.
Sabom, M., & Kreutzinger, S. (1978). Physicians evaluate near death experience. *Theta, 6*, 1–6.
Selleri, F. (1986). *Le grand débat de la théorie quantique*. Paris: Flammarion.
Sokal, A., & Bricmont, J. (1997). *Impostures intellectuelles*. Paris: Odile Jacob.
Suarez, A. (2013). Free will and nonlocality at detection as basic principles of quantum physics. In A. Suarez & P. Adams (Eds.), *Is science compatible with free will?* New York: Springer. Chapter 5.
Van Lommel, P., et al. (2001). Near-death experience in survivors of cardiac arrest: A prospective study in the Netherlands. *The Lancet, 358*, 2039–2045.

Chapter 15
Are Economics Laws Compatible with Free Will?

Luís Cabral

Abstract I argue that determinism at the aggregate level of economic behavior is compatible with uncertainty at the individual level, and that the latter results essentially from individual free will.

Keywords Economics • Free will • Uncertainty

15.1 Introduction

Are economics laws compatible with free will? Such was the question I was challenged with in preparation for the Social Trends Institute (STI) meeting, *Is Science Compatible With Our Desire for Freedom*, Barcelona, October 2010. I am going to be quite frank: this is not a question that economists frequently consider. Or, to put it differently, the great majority of economics researchers will reply that the answer is trivially yes. However, as one digs deeper into the philosophical issues at stake, one realizes that things are not as simple as they seem at first.

In this short paper, I do two things. First, by drawing a parallel with the physics dichotomy of classical mechanics and quantum mechanics, I show how determinism at the aggregate microeconomics level is compatible with uncertainty at the individual microeconomic level. Second, I argue that uncertainty at the individual microeconomic level is observationally consistent with various theories of human behavior, which may or may not make room for free will. Third, I compare the level of predictability (or lack thereof) of behavioral sciences to that of nonbehavioral sciences, arguing that model complexity, rather than human freedom, is the main explanatory factor.

L. Cabral (✉)
W R Berkley Term Professor of Economics, Stern School of Business, New York University

Research Fellow, IME and SPSP (IESE) and CEPR
e-mail: lcabral@stern.nyu.edu

A. Suarez and P. Adams (eds.), *Is Science Compatible with Free Will?: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience*,
DOI 10.1007/978-1-4614-5212-6_15,

Before getting into these three main points, however, I offer for the benefit of readers less acquainted with economics a brief summary of its core elements: the homo economicus model.

15.2 Homo Economicus

Neoclassical economics is largely based on a model of man where rational thought is given to alternatives and decisions are made optimally. This is commonly known as the "homo economicus" model. Mill, J. Stuart, the source of the term, defined the economics approach as follows:

> [Political economy] does not treat the whole of man's nature as modified by the social state, nor of the whole conduct of man in society. It is concerned with him solely as a being who desires to possess wealth, and who is capable of judging the comparative efficacy of means for obtaining that end. (Mill 1836)

Mill and fellow philosopher Bentham, J. went further by proposing the concept of "utility" as the measure of satisfaction that an individual derives from a certain choice or state of the world. The "homo economicus" model could then be rephrased as a process where choices are made in order to maximize utility.

A few decades after Mill proposed the homo economicus model, F. Edgeworth claimed that developments in "physio-psychology" would eventually allow for the direct measurement of the utility mapping. He even coined the term "hedonimeter," the instrument to measure an individual's utility (see Colander 2007).

Edgeworth did not live to see anything close to a "hedonimeter," and it became generally accepted that utility could not be measured or observed directly. The core of neoclassical economics then became largely an axiomatic, deductive process. The basic axiom is that each individual is endowed with a set of preferences, a choice set which is a totally ordered set,[1] and that the individual always chooses the maximum point from that set.

In this context, utility is mainly used as a construct for describing an individual's preferences and actions. In fact, although utility cannot be directly observed, it can be indirectly measured. Based on actual choices made by an individual and based on the above axiom of optimal choice, one can uncover a mapping that gives the value (utility) of each option faced by an individual.[2] This simple idea, which is fairly intuitive though not trivial to prove, was developed by Samuelson and is known as the theory of revealed preference (see Samuelson 1938).

[1] That is, a set of options together with a "preference" binary relation that is transitive, antisymmetric, and total.

[2] Note that this mapping is only unique up to a monotonic transformation. For example, if I say that each of my possible options give me twice as much utility as before, then my optimal choices remain the same.

This is in essence the nature of economics as a behavioral science. Notice that homo economicus is essentially a deterministic model. However, as Stuart Mill points out in the above quote, the homo economicus model is purposely a partial treatment of "man's nature." As such, it would be imprudent to derive conclusions regarding human freedom from the fact that we use a deterministic model. I next turn to this issue in greater detail.

15.3 Heisenberg Uncertainty and Behavioral Uncertainty

Are economics laws compatible with free will? My best answer is to strike an analogy with physics. As M. Heisenberg rightly pointed out,

> Almost 100 years ago, quantum physics eliminated a major obstacle to our understanding of [the issue of freedom and determinism] when it disposed of the idea of a Universe determined in every detail from the outset.

What does this mean in practice? At the scale of planets and at many intermediate scales, quantum effects are of second order, and the deterministic laws of classical mechanics apply reasonably well. In fact, much of what is currently taught at engineering school is still largely drawn from classical mechanics. In other words, there are a series of systems that may safely be treated as "deterministic" even though we know that deep down there is a fundamental source of randomness and unpredictability. More formally, the above statement corresponds to the *correspondence principle*, first formally enunciated in Bohr [1920].

Mutatis mutantis, the same can be said about economics models: at an aggregate level, we may safely treat them as approximately deterministic and "exact" models, even though at the individual level there is a lot of residual randomness. For example, the body "automobile buyers in California" may be treated very much in the same way as a planet would be treated in physics (a body with predictable, deterministic behavior responding to outside influences). However, the body "automobile buyers in California" is composed of individual buyers who, like quantum particles, behave in ways that are at least apparently random.

Let me continue to illustrate the idea with the example of automobile purchases.[3] Suppose there are I individuals ($i = 1, \ldots, I$), each of whom must choose between J different car models ($j = 1, \ldots, J$). A car model may be described by a series of K quantifiable characteristics ($k = 1, \ldots, K$). Examples of car model characteristics might be size, fuel efficiency, acceleration, and so on. Finally, each individual i may be characterized by a series of D demographic indicators $d_{i1},\ldots,d_{iD}$. For example, d_{i1} might be household size or income level.

[3] For a deeper treatment of this type of models, see for example Train [1986].

A typical economic model of consumer choice starts from the notion of individual utility, in our example a measure of how much car model j is worth to consumer i. Suppose that

$$U_{ij} = \sum_{k=1}^{K} \beta_k(d_{i1}, \ldots, d_{iD})\, c_{jk} + \epsilon_{ij} \tag{15.1}$$

In this equation, U_{ij} measures the utility that model j gives consumer i; c_{jk} measures model j's performance along dimension k (e.g., how much car model j possesses the characteristic k, where k may refer to things like size or acceleration—or price, a particularly important characteristic); $\beta_k(d_{i1}, \ldots, d_{iD})$ measures how an individual with demographic characteristics $(d_{i1}, \ldots, d_{iD})$ values performance dimension k. For example, an individual with a large household will value more a larger car, whereas an individual with lower income will value more a car with better fuel efficiency. Finally, ϵ_{ij} measures residual utility of the match between individual i and model j. I will come back to ϵ_{ij} later; in fact, this residual component will be central to my discussion of the relation between economics laws and free will.

As mentioned in the previous section, a basic law of rational economic behavior is that each individual chooses the best option given the information the individual possesses. In the present case, and assuming that each individual knows the values of each car model characteristics c_{jk}, individual i chooses car model j' such that $U_{ij'} \geq U_{ij''}$ for all $j'' \neq j'$.

The formal statement of the above individual-aggregate behavior pattern (random behavior at the "atomic" level and deterministic behavior at the aggregate level) is a statistical convergence theorem, basically the law of large numbers.

Suppose for simplicity that ϵ_{ij} is a random variable with an extreme-value distribution, that is, with c.d.f. $F(\epsilon) = \mathrm{e}^{-\mathrm{e}^{(a-\epsilon)/b}}$.[4] It can be shown that as the number of individuals tends to infinity ($I \to \infty$), then the fraction of individuals choosing j converges almost surely to

$$x_j = \frac{1}{I} \sum_{i=1}^{I} \left(\frac{\exp\left(\sum_{k=1}^{K} \beta_k(d_{i1}, \ldots, d_{iD})\, c_{jk} \right)}{\sum_{\ell=1}^{J} \exp\left(\sum_{k=1}^{K} \beta_k(d_{i1}, \ldots, d_{iD})\, c_{\ell k} \right)} \right)$$

The deterministic nature of the above model can be explained as follows. For a given distribution of demographic characteristics in the population, and for given preference functions β_k, each car's market share is "deterministically" induced by the car's vector of characteristics $\mathbf{c}_j = (c_{jk})$ (including price) vis-a-vis the other cars' vectors of characteristics, $\mathbf{c}_\ell$, where $\ell = 1, \ldots, J$.

[4] This distributional assumption is particularly helpful in that it leads to closed-form expressions for market shares. For other distributions, only numerical solutions can be obtained. Nevertheless, the qualitative points I am making are still valid.

The probabilistic nature of the above model, in turn, can be explained as follows. For a given individual, even if one measures with precision the vector of demographic characteristics $\mathbf{d}_i = (d_{i\ell})$, there is considerable uncertainty as to which car model the consumer will choose. All we can do is to determine the probability of choosing model j, which is given by

$$P_{ij} = \frac{\exp\left(\sum_{k=1}^{K} \beta_k (d_{i1}, \ldots, d_{iD})\, c_{jk}\right)}{\sum_{\ell=1}^{J} \exp\left(\sum_{k=1}^{K} \beta_k (d_{i1}, \ldots, d_{iD})\, c_{\ell k}\right)}$$

15.4 Uncertainty, Measurement Error, and Free Will

There are of course important limitations in my parallel between physics (from quantum mechanics to classical mechanics) and economics (from individual behavior to aggregate behavior). First, the precision of measurement and prediction in the aggregate is considerably weaker in economics than in the physical sciences. In the physical sciences, the values of the relevant parameters can be determined with great precision, almost with arbitrary precision. In economics, by contrast, the values of β_k can only be obtained by statistical estimation based on historical data, a rather poor substitute for laboratory experimentation.

Second, whereas the functional forms that make up a physics model tend to be derived from a coherent conceptual framework, in economics there is a lot of arbitrariness in the choice of particular functional forms. For example, there is nothing in economic theory to indicate that the utility function (15.1) ought to be linear in the βs, or that the error term ϵ distributed according to a particular c.d.f.

More important, the source of randomness in individual behavior in economics is quite different from Heisenberg-type randomness: it is based on individual freedom, not on the behavior of individual particles. Even if I were to know everything about the history of a particular individual, I would still be unable to predict with certainty the individual's behavior when faced with a choice among J alternative options.

This is not an innocuous statement. In fact, uncertainty caused by measurement error and uncertainty caused by genuine free will are to a great extent observationally indistinguishable. In other words, one might argue that the reason I am unable to predict economic behavior is that I really don't know *everything* about the individual's history, possibly including minute details about the individual's brain activity.

Above I mentioned Edgeworth's dream of a "hedonimeter," something he was never able to see in his time. Today, however, we have tools that provide us with an abundance of data regarding the "physio-psychology" process that Edgeworth had in mind. As a result, Neuroeconomics—a combination of psychology, economics,

and neuroscience—has emerged as a recent effort to understand, at the most basic level, how individuals make economics decisions. The idea is to measure and record brain activity at the moment of evaluation and decision making, specifically when an individual must choose between various economic options.[5]

The contribution of neuroeconomics can be thought of at different levels. First, it provides additional individual characteristics that contribute to an individual's choice (i.e.,, it increases the dimensionality of the vector $\mathbf{d}_i$ introduced above). Second, neuroeconomics offers a critique of the economics deliberative model, that is, the fundamental assumption that individuals make rational choices among alternatives. In fact, emotions and automatic responses play a crucial role in human decision making—as psychologists have known for a long time. More important for our present purposes, neuroeconomics provides a "platform" for a theory of deterministic human behavior: the idea that the only source of ϵ uncertainty is measurement error; the idea that if we are able to measure brain activity well enough, then economic behavior will be predictable.

I disagree with this view. I believe there is an irreducible degree of uncertainty which results from each individual's free will. I don't think economists can prove this—and for the reasons described above, I don't think it really matters a great deal from an economics practice point of view.

15.5 Freedom and Predictability

A common complaint faced by economics and economists is that they utterly fail when it comes to predicting events; and a common justification for such failure is that economics deals with "objects" that are endowed with free will (as opposed to nonbehavioral sciences, which deal with nonbehavioral phenomena). I will now try to argue that, notwithstanding fundamental differences between behavioral and nonbehavioral sciences, the "predictability gap" is not as great as many think, and moreover is not primarily caused by the behavioral element, as many argue.

Before talking about the important differences between behavioral and nonbehavioral sciences, it is worth to talk about what they have in common. Scholars who want to understand the world do so by building models, some more formal than others but nevertheless models: conceptual frameworks with various objects, parameters, variables, and relations. Among the great variety of models and realities to be modeled, I find it useful to distinguish between:

(a) Simple models, where a small number of objects and relations are considered; and complex models, where a large number of objects and relations are modeled.[6]

[5] See Glimcher [2003] and references therein.

[6] I am aware that the term "complexity" is frequently used in different senses (as in "complex dynamic systems"), but I could not think of a better term in the present context.

Table 15.1 A taxonomy of models, with examples

	Simple	Complex
Nonbehavioral	Heat and pressure lab experiment	Weather forecasting
Behavioral	Automobile sales	Global economy

(b) Nonbehavioral systems, where no animal or human behavior is involved; and behavioral systems, where either animal or human behavior is involved.

This two-dimensional classification leads to a matrix of possible models, which I illustrate in Table 15.1. In it, I suggest examples for each cell. Take the first row. Any elementary Physics 101 lab experiment would make a good example of something to study with a simple nonbehavioral model. Weather forecasting, by contrast, involves a large number of variables, objects, and relations, though it still deals primarily with nonbehavioral patterns.[7]

Consider now the second row, where I propose examples from economics, one of the leading behavioral fields of study. First, the study of consumer choice of automobile purchases (as described above) provides a good example of a reality to study with a simple model. "Simple" is a relative term: compared to a physics lab experiment there are many more variables to consider in a consumer's purchase decision; but by economies standards this is still a relatively simple decision. Contrast that to analyzing the global economy (e.g., will the world's economies fall into a "double-dip" or will we get out of the current recession? And if so, how soon?). Now we are dealing with a truly complex system, very much like the weather.

My point is that the degree of predictability varies more along the horizontal dimension than it does along the vertical dimension. In other words, it is more difficult to predict the weather than it is to predict automobile purchases.[8] This is not to deny the fundamental difference between nonbehavioral and behavioral models, namely human freedom. Rather, it restates the principle that I mentioned earlier that the law of large numbers provides an analogue in economics to the correspondence principle in physics (roughly, that in the limit quantum mechanics turns into "deterministic" classical physics).

15.6 Concluding Remarks

My main point is that the statistical regularity of aggregate economic behavior is compatible with irreducible uncertainty and unpredictability of individual behavior; and that the latter results from individual free will. In many ways, the point I am

[7] I deliberately avoid models of climate change so as to skirt the issue of human intervention.

[8] As the joke goes, God created meteorologists so that economists looked respectable. (There are, naturally, several versions of this joke, including the one where the roles of economist and meteorologist are reversed).

making about economics can also be made about other human and social sciences. The reason for my particular focus is that, among the social sciences, economics is the field that comes closest to the idea of a deterministic model of the sort offered by classical mechanics.

Finally, the idea that the individual will provide an irreducible source of uncertainty is not exclusive to human behavior. As Heisenberg [2009] points out, animal behavior "cannot be reduced to responses" and thus cannot be predicted based on a deterministic model. But by the same argument developed above, this does not negate the possibility that aggregate animal behavioral is essentially deterministic.

Acknowledgements An earlier draft was presented at the STI meeting, *Is Science Compatible With Our Desire for Free Will*, Barcelona, October 2010. I am grateful to the meeting's participants, Bob Doyle in particular, for comments and suggestions. Alas, I remain the sole responsible for errors and shortcomings.

References

Bohr, N. (1920). Über die Serienspektra der Element. *Zeitschrift für Physik*, *2*, 423–478.

Colander, D. (2007). Retrospectives: Edgeworth's Hedonimeter and the quest to measure utility. *Journal of Economic Perspectives*, *21*, 215–226.

Glimcher, P. (2003). *Decisions, uncertainty, and the brain: The science of neuroeconomics*. Cambridge: MIT.

Heisenberg, M. (2009). Is free will an illusion? *Nature*, *459*, 164–165.

Mill, J.S. (1836). On the definition of political economy, and on the method of investigation proper to it. *London and Westminster Review*, October.

Samuelson, P. (1938). A note on the pure theory of consumers' behaviour. *Econometrica*, *5*, 353–354.

Train, K. (1986). *Qualitative choice analysis: Theory, econometrics, and an application to automobile demand*. Cambridge: MIT.

Part III
Attempts to Reconcile Science and Free Will

Chapter 16
The Two-Stage Model to the Problem of Free Will

How Behavioral Freedom in Lower Animals Has Evolved to Become Free Will in Humans and Higher Animals

Robert O. Doyle

Abstract Random noise in the neurobiology of animals allows for the generation of alternative possibilities for action. In lower animals, this shows up as behavioral freedom. Animals are not causally predetermined by prior events going back in a causal chain to the origin of the universe. In higher animals, randomness can be consciously invoked to generate surprising new behaviors. In humans, creative new ideas can be critically evaluated and deliberated. On reflection, options can be rejected and sent back for "second thoughts" before a final responsible decision and action.

When the indeterminism is limited to the early stage of a mental decision, the later decision itself can be described as adequately determined. This is called the two-stage model, first the "free" generation of ideas, then an adequately determinism evaluation and selection process we call "will."

Keywords Free will • Determinism • Two-stage model • Chance • Randomness

16.1 Introduction

In the May 14, 2009 issue of Nature Magazine, Heisenberg Martin (Heisenberg 2009) challenged the idea, popular in the recent psychology and philosophy literature, that human free will is an free will illusion (Wegner 2002). Heisenberg suggested that a lot could be learned by looking at animals, to see how they initiate behavior. The behaviorist idea that actions are deterministic causal responses to external stimuli has been discredited. For decades, Watson–Skinner behaviorism focused on stimulus and response. They ignored the existence of internal states in

R.O. Doyle (✉)
Astronomy Department, Harvard University, Cambridge, MA, USA
e-mail: rodoyle@fas.harvard.edu; bobdoyle@informationphilosopher.com

A. Suarez and P. Adams (eds.), *Is Science Compatible with Free Will?: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience*,
DOI 10.1007/978-1-4614-5212-6_16,

the mind, but today such internal mental states are accepted as the causes of actions, in animals and humans. Can these mental states themselves be only statistically "caused?" Can mental states—thoughts and ideas—involve an indeterminism which breaks the deterministic causal chain to all events in the "fixed" past, but which does not make our actions themselves "random?"

In my own correspondence with Nature in their June 25, 2009 issue (Doyle 2009), I connected Heisenberg's thinking with James, William's 1884 two-stage model of free will (James 1956, p. 145). The first stage is the chance generation of possibilities, alternative (ideas just "pop into our heads"). The second stage is a "willed" decision "caused" by our reasons, motives, and feelings that help an agent evaluate and "select" among the first-stage alternative possibilities. In the second stage, the agent evaluates the options in a "determined" way, but not one that was "predetermined" from the time before the new possibilities were generated (Doyle 2010).

Long before twentieth-century behaviorism and logical empiricism had limited the study of the mind to externally observable phenomena, James had argued in *The Dilemma of Determinism*, that random chance played a role in generating alternative possibilities.

> The stronghold of the determinist argument is the antipathy to the idea of chance. . .This notion of alternative possibility, this admission that any one of several things may come to pass is, after all, only a roundabout name for chance. (James 1956, p. 153, Doyle 2010)

And James explicitly connected spontaneous variations in the evolution gene pool with random images and thoughts in the human brain.

> [In mental evolution], if anywhere, it would seem at first sight as if that school must be right which makes the mind passively plastic, and the environment actively productive of the form and order of its conceptions; which, in a word, thinks that all mental progress must result from a series of adaptive changes, in the sense already defined of that word. . .It might, accordingly, seem as if there were no room for any agency other than this; as if the distinction we have found so useful between "spontaneous variation," as the producer of changed forms, and the environment, as their preserver and destroyer, did not hold in the case of mental progress; as if, in a word, the parallel with Darwinism might no longer obtain. . .And I can easily show. . .that as a matter of fact the new conceptions, emotions, and active tendencies which evolve are originally produced in the shape of random images, fancies, accidental out-births of spontaneous variation in the functional activity of the excessively instable human brain. (James 1880)

Heisenberg, Martin thus became the latest in a long list of philosophers and scientists who sought a "two-stage" model (see http://informationphilosopher.com/freedom/two-stage_models.html), a temporal sequence of first acausal randomness, then causal law-like selection, as the basis for human freedom. Before Heisenberg, the question always was how to free the *human* brain from deterministic worries. Now that Heisenberg has extended the concept of randomly generated alternative possibilities for action throughout the animal kingdom, he has liberated all life from the complete predeterminism implied by the Newtonian and Laplacian world view of William James's time.

16.2 Antipathy to Chance and the Standard Argument against free will

What James, William called the "antipathy to chance" goes back 2,300 years to the Stoic and Academic philosophers' attack on Epicurus' notion of an atomic "swerve." Epicurus said such a random swerve was needed to break the bonds of his materialist and atomist colleague Democritus, whose strict causal physical determinism denied human freedom (Lucretius 1982). Stoics and Academics attacked Epicurus for suggesting that human freedom was the result of chance. That, they said, would make our actions random and deny human responsibility (Cicero 1951). For the Stoics, Nature was identical to God and Reason (Long 1986). To suggest that chance really exists in Nature invites the atheistic thought that God is either irrational or ignorant of future events.

The standard argument *against* free will is the very simple and logical claim that either determinism or indeterminism is true. If determinism is "true," we are not free, if indeterminism is "true," we are not responsible (Ayer 1954; Doyle 2011, Chap. 4).

Our free-will model of two stages in a temporal sequence is motivated by the need to answer the two objections to free will in the standard argument against it. Limiting indeterminism to the first stage prevents it from making our decisions themselves *random*, which would threaten our responsibility. The "determinism adequate" of the second stage defeats the problem of *predeterminism* from the Big Bang that threatens our freedom. By "adequate" determinism we mean that there may be some low level of indeterminism in the second stage but it is statistically irrelevant.

In the logical choice between the "truth" of determinism or indeterminism, it is indeterminism that is "true" in the universe, but many microscopic random events are averaged over and irrelevant in the macroscopic world. Nevertheless, most philosophers today are determinist and compatibilist, unless they embrace a metaphysical dualism (Swinburne 2011). And many scientists claim that the brain is determined (cf. Gazzaniga 2011).

We can see why so many philosophers accept the idea that determinism is "compatibilism" with free will. It is because given the forced choice between the determinism and indeterminism in the standard argument, determinism at least makes our actions responsive to reasons. They can be caused by our motives, feelings, and desires. They result from a nonrandom deliberation that evaluates our options.

What Heisenberg, Martin and many other thinkers have established is that randomness at some level or stage (the generation of alternative possibilities) need not jeopardize adequate law-like behavior at another level or stage (the adequately determined evaluation of those possibilities).

As long ago as 1690, Locke, John insisted on the separation of "free" and "will." He hoped

> to put an end to that long agitated, and, I think, unreasonable, because unintelligible, question, viz. *Whether man's will be free or no?* For if I mistake not, it follows from what I have said, that the question itself is altogether improper. . .This way of talking, nevertheless, has prevailed, and, as I guess, produced great confusion. . .I think the question is not proper, *whether the will be free*, but *whether a man be free*. (Locke 1959) [Locke's emphasis.]

A century later, Hume, David "reconciled" man's freedom with determinism in the notion we now call "compatibilism." He properly insisted that our will is determined by our motives and inclinations.

> to proceed in this reconciling project with regard to the question of liberty and necessity; the most contentious question of metaphysics, the most contentious science; it will not require many words to prove, that all mankind have ever agreed in the doctrine of liberty as well as in that of necessity, and that the whole dispute, in this respect also, has been hitherto merely verbal.
>
> By liberty, then, we can only mean *a power of acting or not acting, according to the determinations of the will*; this is, if we choose to remain at rest, we may; if we choose to move, we also may. Now this hypothetical liberty is universally allowed to belong to every one who is not a prisoner and in chains. Here, then, is no subject of dispute. (Hume 1975, p. 95)

But Hume denied that liberty depended on chance. For Hume and the great mathematicians who developed the calculus of probabilities—Abraham de Moivre before Hume and Laplace, Pierre-Simon after him, chance was merely human ignorance.

> liberty, when opposed to necessity, not to constraint, is the same thing with chance; which is universally allowed to have no existence. (Hume 1975, p. 56)
>
> Though there be no such thing as *Chance* in the world; our ignorance of the real cause of any event has the same influence on the understanding, and begets a like species of belief or opinion. (Hume 1975, p. 96)

Nevertheless, Hume recognized a serious objection to his theory, that everything might be predeterminism. Most compatibilists and determinists since Hobbes and Hume never mention the fact that a causal chain of events going back before our birth would not provide the kind of liberty they are looking for. But Hume frankly admits that such a causal chain would be a serious objection to his theory.

> I pretend not to have obviated or removed all objections to this theory, with regard to necessity and liberty. I can foresee other objections, derived from topics which have not here been treated of. It may be said, for instance, that, if voluntary actions be subjected to the same laws of necessity with the operations of matter, there is a continued chain of necessary causes, pre-ordained and pre-determined, reaching from the original cause of all to every single volition, of every human creature. No contingency anywhere in the universe; no indifference; no liberty. While we act, we are, at the same time, acted upon. (Hume 1975, p. 99)

Today we can finally reconcile free will with chance, randomness, and Indeterminism, which alone can break this "continued chain of necessary causes."

16.3 Chance and Randomness in Cosmology and Biology

Randomness has been present in cosmology since the origin of the universe, a state of total chaos (minimal information) nearly 14 billion years ago. But mathematicians and physicists sought deterministic explanations that attempt to avoid randomness. The most famous was Pierre-Simon Laplace, who in 1815 postulated a super-intelligence that could know the positions, velocities, and forces on all the particles

in the universe at one time, and thus know the universe for all past and future times. This implies that information is a constant of nature. Some mathematicians think that information is a conserved quantity—like matter and energy.

But midway through the nineteenth century, Kelvin, Lord (William Thomson) realized that the newly discovered second law of thermodynamics required that information could not be constant, but would be destroyed as the entropy (disorder) increased. Hermann Helmholtz described this as the heat death of the universe.

Kelvin's claim would be correct if the universe were a closed system. But in our open and expanding universe, Layzer, David showed that the maximum possible entropy is increasing faster than the actual entropy (Layzer 1975). The difference between maximum possible entropy and the current entropy is called negative entropy, opening the possibility for complex and stable information structures.

Despite the second law of thermodynamics, stable and law-like information structures evolved out of the chaos, first, in the form of microscopic particulate matter—quarks, baryons, nuclei, and electrons, then later, under the influence of gravitation—macroscopic galaxies, stars, and planets. Every new Information structure reduces the entropy locally, so the second law requires an equal (or generally much greater) amount of entropy to be carried away. Without the expansion of the universe, this would be impossible.

Whether the newly formed stable structure is a baryon or a planet, the new "bits" of information can be regarded as physical "measurements" that involve the collapse of quantum mechanical wave functions. Ludwig, Gunter (Ludwig 1953) and Landauer, Rolf (Landauer 1961) showed that any such measurement that increases the number of information bits must involve a compensating increase in the entropy or randomness elsewhere. For Ludwig, it was in the measurement apparatus. For Landauer, it was the energy dissipated by a computer's power supplies.

Because of the "Law of Large Numbers" in statistics, and the Correspondence Principle of Quantum mechanics (which says that quantum physics approaches classical physics for large quantum numbers), the Newtonian laws of classical mechanics, discovered in the stable and regular motions of the planetary orbits, are "Determinism, Adequate. Events are normally determined by immediate prior events, but not strictly *pre*determinism from the origin of the universe. This is despite the residue of real originary chaos in many parts of the universe, especially in the quantum-mechanical microcosmos. The effects of Quantum Indeterminacy can thus normally be ignored in the macroscopic world of classical physics. (The second stage of Two-stage Model assumes that microscopic indeterminacy can be ignored in the evaluation/selection stage.)

Whereas randomness can normally be ignored in macroscopic physics, randomness in biology plays a central role, in the evolution of species and in the life strategies of many organisms, not only animals. Darwin was circumspect and cautious about "mere chance," because in his time chance still evoked strong atheistic sentiments.

In animals, Heisenberg, Martin cites the bacterium *Escherichia coli* (Heisenberg 2009, p. 165). These tiny organisms are equipped with sensors and motion

capability that let them make two-stage decisions about which way to go. They can move in the direction of nutrients and away from toxic chemicals. They do this with tiny flagella in their tails that rotate in two directions. Flagella rotating clockwise cause the bacterium to tumble and face random new directions. When the flagella rotate counter-clockwise, the bacterium moves forward and sensory receptors on the bacterium surface detect gradients of chemicals and temperatures. If the gradient indicates "food ahead," or perhaps "danger behind," the bacterium continues straight ahead. The law-like decision to go forward is an adequately determined evaluation of sensors along the bacterium's body. If the sensed gradients are unsuitable, the flagella reverse and the bacterium again tumbles.

We see that even the lowest forms of animal can recruit randomness to serve their teleonomic purposes. Mayr, Ernst has shown that evolution is conservative, reusing existing mechanisms rather than inventing new ones. So what Mayr calls the "two-step" process (Mayr 1988) of Darwinian evolution itself may have become a feature of living organisms up to higher animals and humans.

The mind's "two-stage" ability to be creative and free is likely evolved *indirectly* from Mayr's "two-step" process and then *directly* from the combination of random and law-like behavior in the lower animals. Free will is therefore not an *ad hoc* development in humans, as many philosophers (especially theologians) have thought. It is a normal biological property, not a gift of God or an inexplicable mystery. We may not have metaphysical free will, but we do have biophysical Behavioral Freedom. Our lives are not predeterminism.

16.4 Four Evolving Selection Levels

The development path from behavioral freedom in the lower animals to free will in humans has primarily involved significant changes in the complexity of the second stage—the evaluation and selection process.

Randomness in the first stage always has the same source—namely chaotic thermal and Quantum noise. It is the second-stage selection process itself that has significantly evolved. We can identify different levels of selection, but note that at each level organisms use all the earlier types of selection as well.

Natural selection—for biological evolution, selection is reproductive success for a population.

Instinctive selection—by animals with little or no learning capability. Selection criteria are transmitted genetically.

Learned selection—for animals whose past experiences guide current choices. Selection criteria are acquired environmentally, including instruction by parents and peers.

Predictive selection—using imagination and foresight to evaluate the future consequences of choices.

Reflective and normative selection—in which conscious deliberation about cultural values influences the choice of behaviors.

Evolution has added more and more features to selection over time, instinct, learning, prediction, and reflection. These eventually become the many factors at work in the fully conscious human will.

16.5 Randomness in Psychology and Philosophy

Real (ontological, not epistemological) chance was welcomed by at least one philosopher and psychologist of the nineteenth century, namely James, William. But since the twentieth-century discovery of real chance in the form of quantum indeterminacy by Heisenberg, Werner, chance and randomness have not fared well in psychology or philosophy.

In his Gifford Lecture of 1927, Eddington, Arthur Stanley had described himself as unable "to form a satisfactory conception of any kind of law or causal sequence which shall be other than deterministic." (Eddington 1958). Yet just a year later, in response to Heisenberg's indeterminacy principle, Eddington revised his lectures for publication as *The Nature of the Physical World*. There he dramatically announced, "It is a consequence of the advent of the quantum theory that *physics is no longer pledged to a scheme of Determinism law*" (Eddington 1958, p. 295). He went even farther and enthusiastically identified indeterminism with freedom of the will.

But the critical reaction of philosophers was swift (see Stebbings 1958). A "free electron" has nothing to do with "free will," they complained. A Brain, quantum event in, amplified to affect our reasoning, can only make our decisions random. Quantum events simply happen to us. They are not "up to us." We are not responsible for them. Late in life, Eddington yielded to the criticism, saying that he could find no "half-way house" between determinism and indeterminism (Eddington 1938).

["Up to us" or "depends on us" (ἐφ' ἡμῖν) was for the Greeks, and particularly for Aristotle, the term closest to the modern complex idea of free will (which combines freedom and determination in an apparent internal contradiction). Aristotle and Epicurus both said something "up to us" was a "third thing" that was neither chance nor necessity. The idea was a kind of "agent causality" that provides accountability or moral responsibility. Because our actions originate "within ourselves" (ἐν ἡμῖν), they say that as "agents" we are "causes."]

A number of prominent philosophers and scientists struggled to include quantum indeterminacy in a model of free will, including Compton, Arthur Holly (Compton 1931), Margenau, Henry (Margenau 1968), and Popper, Karl (Popper 1977). But their efforts were not convincing to the philosophical community and are rarely referenced in the free will debates.

The one living philosopher who has spent his adult career trying to explain free will as involving quantum events is Kane, Robert. Kane has had some significant success showing that we can be *Responsibility* for an event even if it happens

indeterministically. He considers the case of a businesswoman on the way to an important meeting when she observes an assault in an alley (Kane 1999). She has excellent (moral and humanitarian) reasons to help the victim. She has equally important (practical and self-interested) reasons to continue on and advance her career.

Kane argues that whichever way the businesswoman decides, and even if the "torn decision," as he calls it, is undetermined as a result of neural noise, she has excellent reasons to take responsibility either way. But Kane himself has not found two-stage free will models everything that is needed (Kane 2005), and other prominent libertarian philosophers like van Inwagen, Peter have said that "free will remains a mystery (van Inwagen 2000)."

Some philosophers have been critical of Kane and argue that the agent cannot claim responsibility if the decision was at all random and thus a matter of "luck." The idea of "Luck, Moral" is the source of many moral paradoxes and dilemmas (Nagel 1979; Williams 1981). If something happens entirely by luck, good or bad luck, it appears to be not our responsibility. But Kane's solution to the problem of an indeterministic decision between multiple alternatives, each supported by excellent reasons and motives, solves this problem of luck. The agent can take full responsibility, however she decides. And the specific "cause" of the resulting action is the excellent reason she has for doing it, says Kane.

Mele, Alfred considered a two-stage model of free will in which indeterminism (he called it incompatibilism) is confined to the early stage (Mele 1995). The latter stage he describes as "compatibilist" (effectively and adequately determined). Mele's model is similar to one proposed much earlier by Dennett, Daniel (Dennett 1978). Dennett's work incorporated the still earlier ideas of Wiggins, David (Wiggins 1973), Popper, Karl, and Compton, Arthur Holly.

Dennett did not endorse his own two-stage decision model because he could not imagine a plausible location for quantum events in the brain, one exquisitely timed to be of help in the decision process. How could a randomly timed event be of any help? He settled instead for pseudo-random number sequences (like those generated by a completely deterministic computer program) as all that is needed in his decision-making model.

In a recent book, Mele considered the problem of free will and luck (Mele 2006), comparing the indeterministic early stage of his model to a neural roulette wheel in the head, with a tiny neural ball whose probabilities may be high for landing in the wheel segment for action A, but it is still luck that it did not land in the segment for action B. In the end Mele, like Dennett, could not endorse a two-stage model.

16.6 The Basic Freedom, Requirements for Human

Freedom requires the randomness of absolute chance to break the causal chain of determinism (actually predeterminism), yet it must provide the conscious knowledge that we are adequately determined to be responsible for our choices, that our decisions and actions are "up to us."

Freedom requires some events that are not causally determined by immediately preceding events, events that are unpredictable by any agency, events involving quantum indeterminacy.

These random events can generate alternative possibilities for action. They are the source of the creativity that adds new information to the universe. Randomness is the "free" in free will.

Freedom also requires an adequately determined will that chooses or selects from those alternative possibilities. There is effectively nothing uncertain about this choice. "Adequate" determinism is the determination, the "will" in free will.

Determinism, Adequate means that randomness in our thoughts about alternative possibilities does not directly cause our actions.

Random thoughts can therefore lead to intentions, evaluations, and decisions that are adequately determined to produce actions, for which we can take moral responsibility.

Thoughts *come to us* freely. Actions *go from us* willfully.

We must *admit indeterminism,* but not *permit it* to produce random actions as determinists mistakenly fear.

We must also *limit determinism,* but not *eliminate it* as libertarians mistakenly think necessary.

Evaluation and careful deliberation of all the available possibilities, both ingrained habits and creative new ideas, must help us to "determine" and thus "cause" our actions.

But event *acausality* somewhere is a prerequisite for any kind of agent *causality* that is not *predetermined.*

We thus define "free will" as a two-stage creative process in which a human or higher animal freely generates alternative possibilities, some caused by prior events, some uncaused, following which the possibilities are evaluated and one is "willed," i.e., selected or chosen for adequately determined reasons, motives, or desires.

16.7 How Quantum Noise Can Help Free Will and Not Compromise Responsibility

In my two-stage model of free will and creativity, randomness is not (normally) the direct cause of our actions, but rather simply the free generator of Possibilities, alternative for the Determinism, Adequate will to evaluate and select. I call this noisy generator of creative ideas the "Micro Mind."

An important additional requirement is that the adequately determined will, which I call the "Macro Mind," must have the power to invoke the generation of alternative possibilities (turn it on when needed and off when it is simply interfering with thought processes). For example, the bacterium in Heisenberg's example can turn on randomness by reversing the direction of flagella rotation. This is sometimes called "downward causation (Murphy et al. 2009)." It is not that the mind is

actually controlling specific quantum events. Quantum events are uncontrollable. But the mind can turn access to quantum randomness off, and on again when chance is needed to produce new ideas.

The Micro Mind is different from the early stage in previous two-stage models because it does not depend on a *single quantum event* in the brain that gets amplified to the Macro Mind. The insoluble problem for previous two-stage models has been to explain how a random event in the brain can be timed and located —perfectly synchronized!—so as to be relevant to a specific decision. The answer is it cannot be, for the simple reason that quantum events are totally unpredictable. The mind, like all biological systems, has evolved in the presence of constant noise and is able to ignore that noise when it is unhelpful. It can utilize that noise when it provides a significant competitive advantage, which it clearly does as the basis for freedom and creativity in the first stage of my two-stage model.

Rather than search for a single cause behind a decision, we assume that there are always many contributing causes for any event, and in particular for a mental decision. The two-stage model does not depend on single random events, one per decision. It recruits many random events in the brain as a result of ever-present *noise*, both quantum and thermal noise, that is inherent in any information storage and communication system.

In the Newell-Simon "Blackboard" mind model (Newell and Simon 1972) and Bernard Baars' "Theater of Consciousness" and "Global Workspace" models (Baars 1997), there are always many competing possibilities for our next thought or action. Some of these possibilities may be traceable to causal chains that we ourselves did not initiate. Many possibilities are the result of genetic inheritance or environmental conditioning, for example. Some are well-established habits that are the result of what Robert Kane calls "self-forming actions" (Kane 1984) that happened long ago.

Each of these possibilities is the result of a sequence of events that goes back in an assumed causal chain until its beginning in an uncaused event.

If we could trace any particular sequence of events back in time, it would come to one event whose major contributing cause (or causes) was itself uncaused (a *causa sui*).

For Aristotle, every series of causes "goes back to some starting-point (ἀρχή), which does not go back to something else. This, therefore, will be the starting-point of the fortuitous, and nothing else is the cause of its generation." (Aristotle 1933a)

We can thus in principle assign times, or ages, to the starting points of the contributing causes of a decision. Some of these may in fact go back before the birth of an agent, hereditary causes for example. To the extent that such causes adequately determine an action, we can understand why hard determinists think that the agent has no control over such actions. (Of course, if we can opt out of a habitual action at the last moment, we retain a kind of control. We can always just say no!)

Other contributing causes may be traceable back to environmental and developmental events, perhaps education, perhaps simply life experiences, that were "character-forming" events. These and hereditary causes would be present in the

mind of the agent as fixed habits, with a very high probability of "adequately determining" the agent's actions in many well-understood situations.

But other contributing causes of a specific action may have been undetermined up to the very near past, even fractions of a second before an important decision and moments after the "circumstances" mistakenly thought by some compatibilists to *determine* the action. The causal chains for these contributing causes originate in the noisy brain. They include the free generation of new alternative possibilities for thought or action during the agent's deliberations. They fit Aristotle's criteria for causes that "depend on us" (ἐφ' ἡμῖν) and originate "within us" (ἐν ἡμῖν). (Aristotle 1933b)

Causes with these most recent starting points are the fundamental reason why an agent can *do otherwise* in what are essentially the *same circumstances* (up to the starting point of considering options).

These alternatives are likely generated from our internal knowledge of practical possibilities based on our past experience. Those that are handed up for consideration to Baars' "executive function" may be filtered to some extent by unconscious processes to be "within reason." They likely consist of random variations of past actions we have willed many times in the past.

Note that the random events that generate a new possibility need not be located in the brain itself, nor even be contemporaneous with the immediate decision. It could have been an idea first generated years ago and only now acted upon. And it could have had its origin external to the brain, in the ideas of other persons or in environmental accidents. It need only "come to mind" during deliberations, which itself is partly a matter of luck. But as with the "problem of luck" discussion above, the chance element in the first stage does not make the second-stage decision itself random.

Note also that the evaluation and selection of one of these possibilities by the will in the second stage is as deterministic and causal a process as anything that a determinist or compatibilist could ask for, consistent with our current knowledge of the physical world.

But remember that instead of strict causal determinism, the second stage offers only *adequate* determinism. The random origins of possibilities in the first stage provide freedom of thought and action. As long as the Micro Mind can create new alternative possibilities, we can be free.

16.8 A More Detailed Look at the Micro Mind

Imagine a Micro Mind with a randomly assembled "agenda" of possible things to say or to do. These are drawn from our memory of past thoughts and actions, but randomly varied by unpredictable negations, associations of a part of one idea with a part or all of another, and by substitutions of words, images, feelings, and actions drawn from our experience. In information communication terms, there is cross-talk and noise in our neural circuitry.

In a "content-addressable" information model, memories are stored based on their content—typically bundles of simultaneous images, sounds, smells, feelings, etc. So a new experience is likely to be stored in neural pathways alongside closely related past experiences. A fresh experience, or active thinking about an experience that presents a decision problem, is likely to activate nearby brain circuits, ones that have strong associations with our current circumstances. These are likely to begin firing randomly, to provide unpredictable raw material for actionable possibilities.

The strong feeling that sometimes "we don't know what we think until we hear what we say" reflects our capability for original and creative thoughts, different from anything we have consciously learned or thought before. A new idea may be something as simple as substituting a synonymous word, or more complex replacements with associated words (metonyms) or wild leaps of fancy (metaphor) are examples of building unpredictable thoughts. Picturing ourselves doing something we have seen others do, from "monkey see, monkey do" childhood mimicry to adult imitations, is a source for action items on the agenda, with the random element as simple as if and when we choose to do them.

But how exactly is the required randomness recruited to build these alternative possible thoughts and actions?

Some critics argue that brain structures are too large to be affected at all by quantum events. But there is little doubt that the brain has evolved to the point where it can access quantum phenomena. The evolutionary advantage for the mind is freedom and creativity. Biophysics tells us the eye can detect a single quantum of light (a photon), and the nose can smell a single molecule. It seems clear that the brain has evolved to the quantum limit and thus has access to quantum noise—when randomness is helpful, when it enhances reproductive success.

If the Micro Mind is a random generator of frequently outlandish and absurd possibilities, the complementary Macro Mind is a macroscopic structure so large that quantum effects are negligible. It is the critical apparatus that makes adequately determined decisions based on our character and values. It can suppress quantum noise, by averaging over many such effects to achieve statistical regularity, or perhaps even with the kinds of error detection and correction techniques designed into modern computers.

Note that information about our character and values is probably stored in the same noise-susceptible neural circuits of our brain cortex. Macro Mind and Micro Mind are not necessarily in different locations in the brain. Instead, their difference is probably the consequence of different information processing methods. The Macro Mind must suppress the noise when it makes an adequately determined decision. But it also can turn on the sensitivity to noise in the Micro Mind when new possibilities are needed.

Normally noise is the enemy of information, but it can be the friend of freedom and creativity.

The Macro Mind has very likely evolved to add enough redundancy to reduce the noise to levels required for an adequate determinism. This means that our decisions are in principle predictable, given knowledge of all our past actions and given the randomly generated possibilities in the instant before decision. However,

only we know the contents of our minds. New possibilities exist only within our minds. So other persons could not predict our actions, and until neuroscientists can resolve the finest details of information storage in our brains, they too could not predict our thoughts and decisions.

The two-stage model accounts not just for freedom but also for creativity, original thoughts and ideas never before expressed. Unique and new information may come into the world with each new thought and action. We are the originators of the new information, the authors of our lives, and in this respect we are co-creators of our universe (Doyle 2011; Chap. 22).

Biologists will note that the Micro Mind corresponds to random variation (mutations) in the gene pool (often the direct result of quantum accidents). The Macro Mind corresponds to natural selection by highly determined organisms. Karl Popper may have been the first to point this out (Popper 1977).

Psychologists will see the resemblance of Micro Mind and Macro Mind to the Freudian id and super-ego (*das Es und das Über-ich*).

Note that the two-stage model accounts quantitatively for the concept of wisdom. The greater the amount of knowledge and experience, the more likely that the random Agenda will contain more useful and "intelligent" thoughts and actions as alternative possibilities. It also implies that an educated mind is "more free" because it can generate a wider Agenda and options for action. It suggests that "narrow" and "closed" minds may simply be lacking the capabilities for generating new ideas of the Micro Mind. And if the Macro Mind were weak, it might point to the high correlation between creativity and madness suggested by a Micro Mind out of control, or it might be an indicator for Aristotle's "weakness of will" (*akrasia*).

Philosophers of mind, whether determinist or compatibilist, should recognize that the second-stage Macro Mind has everything they say is needed to make a carefully reasoned and responsible free choice. But now our choices include self-generated random possibilities for thought and action that no external agent can predict. Thus the choice of the will and the resulting willed action are unpredictable. The origin of the chosen causal chain is entirely within the agent, a condition noted first by Aristotle for voluntary action, the causes are "in us" (ἐν ἡμῖν). The two-stage model clearly describes "self-determination."

16.9 Decisions are a Multistep, Even Continuous, Process

The two-stage model is not limited to a single stage of generating alternative possibilities followed by a single stage of determination by the will.

It is better understood as a continuous process of possibilities generation by the Micro Mind (the parts of the brain that leave themselves open to noise) and adequately determined choices made from time to time by the Macro Mind (the same brain parts, perhaps, but now averaging over and filtering out the noise that might otherwise make the determination random).

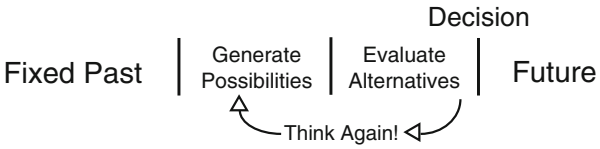


Fig. 16.1 The two-stage model of free will

In particular, note that a special kind of decision might occur when the Macro Mind finds that none of the current options are good enough for the agent's character and values to approve. The Macro Mind then might figuratively say to the Micro Mind, "Think again!"

Many philosophers have puzzled how an agent could do otherwise in *exactly* the same circumstances. Given the myriad of possible circumstances, it is impossible that an agent is ever in *exactly* the same circumstances. The agent's memory (stored information) of earlier similar circumstances guarantees that.

But given the "laws of nature" and the "fixed past" just before a decision, philosophers wonder how a free agent can have any possible alternatives. This is partly because they imagine a timeline for the decision that shrinks the decision process to a single moment.

Collapsing the decision to a single moment between the closed fixed past and the open ambiguous future makes it difficult to see the free thoughts of the mind followed by the willed and adequately determined action of the mind and body.

The view of two stages in a temporal sequence makes a somewhat artificial separation between the creative randomness of the Micro Mind and the deliberative evaluation of the Macro Mind. These two capabilities of the mind can be going on at the same time. As Fig.16.1 shows, this can be visualized by the occasional decision to go back and think again, when the available alternatives are not good enough to satisfy the demands of the agent's character and values.

Our thoughts are free and often appear to come to us. Our actions are adequately determined for moral responsibility and appear to come from us. They are up to us (Aristotle's ἐφ' ἡμῖν).

What then are the sources of alternative possibilities? To what extent are they our creations? We can distinguish three important sources, all of them capable of producing indeterministic options for thoughts and actions. Two come in from outside the mind, the third is internal.

The first source is the external world that arrives through our perceptions. It is perhaps the major driving force in our lives, constantly requiring our conscious attention. Indeed, consciousness can be understood in large part as the exchange of actionable information between organism and environment. Although the indeterministic origin of such ideas is outside us, we can take full responsibility for them if they become one of our adequately determined willed actions.

The second source of options is other persons. The unique human ability to communicate information means that alternative possibilities for our actions are being generated by our reactions to other minds.

Finally, and most importantly, the Micro Mind generates possibilities *internally*. Alternative possibilities truly originate within us (Aristotle's ἐν ἡμῖν). In the two-stage model, the agent is a creative source, the author and originator of her ideas.

16.10 Six Ways Chance Contributes to Free Will

1. Chance exists in the universe. Quantum mechanics is correct. Indeterminism is true, etc.
2. Chance is necessary for free will. It breaks the causal chain of predeterminism.
3. Chance does not directly cause our actions. We can only be responsible for random actions if we flip a coin and claim responsibility "either way."
4. Chance can only generate random (unpredictable) alternative possibilities for action or thought. The choice or selection of one action must be adequately determined, so that we can take responsibility. And once we choose, the connection between mind/brain and motor control must be adequately determined to see that "our will be done."
5. Chance, in the form of noise, both quantum and thermal, must be ever present. The naive model of a single random microscopic quantum event, amplified to affect the macroscopic brain, never made sense. Under what *ad hoc* circumstances, at what time, at what place in the brain, would it occur to influence a decision?
6. Chance must be overcome or suppressed by the adequately determined will when it decides to act, de-liberating the prior free options that "one could have done."

Earlier two-stage models have embraced the first two of these roles for chance, but very few thinkers, if any, appear to have considered all six essential requirements for chance to contribute to libertarian free will.

16.11 How Does the Two-Stage Model Improve on Other Recent Free-Will Views?

The two-stage model lies *between* the work of libertarians and compatibilists, who believe that free will is compatible with determinism.

Apart from religious thinkers, who think free will is a gift of God, and metaphysical dualists, who think freedom lies in an immaterial noumenal realm, the leading libertarian model is that of Robert Kane and his followers Laura Waddell Ekstrom (Ekstrom 2000) and Mark Balaguer (Balaguer 2010). They and Kane's critic Richard Double (Double 1991) have all reached for the dream of genuine indeterminacy "centered" in the "moment of choice," while nevertheless achieving agential control over actions.

Kane calls it "dual voluntary control" when an agent has good reasons for deciding either way in a "torn" decision. So the choice can be random and yet the agent still can feel responsible. We accept Kane's clever argument for responsibility "either way." But it seems confusing to describe this as "control" at the moment of choice when the final choice is avowedly random, and Kane's critics have strongly objected.

Double started out trying to justify three Kane conditions for free will—control, rationality, and dual/plural alternative possibilities that allow the agent to choose otherwise in exactly the same circumstances.

But in the end Double concluded that these three conditions could not be met simultaneously by Kane's model and said so in his 1990 book The Non-Reality of Free Will. To be sure, Double may simply share the goal of "Impossibilists" like Galen Strawson (Strawson 1994), or "Hard Incompatibilists" like Derk Pereboom (Pereboom 2001) or "Illusionists" like Saul Smilansky (Smilansky 2000). All these thinkers share a goal. They want to deny moral responsibility in order to eliminate moral "desert" and retributive punishment. But responsibility can be separated from punishment (see http://www.informationphilosopher.com/freedom/separability.html).

Let's see how my two-stage model can improve on Kane's example of the businesswoman mentioned above. Recall that she is "torn" between helping the victim in the alley and continuing to her important business meeting. Before she decides (randomly) between the given choices, she can activate her alternative possibilities generator and the Micro Mind might come up with additional alternative possibilities. She might for example continue on to her meeting but get out her cell phone to report the crime and call for assistance. On her way she might tell any passersby to go to the victim's aid. Note that these creative new options can "come to her" up to and even beyond the moment of choice in this case (she is on her way to the office).

So my two-stage model with the generation of alternative possibilities appears to provide real freedom beyond earlier two-stage models that Kane properly found unacceptable.

The leading thinkers to have proposed but not endorsed a two-stage model are the compatibilist Daniel Dennett (Dennett 1978, p. 286) and the agnostic Albert Mele (Mele 1995, p. 212). Neither of them could see how quantum events could provide an intelligible explanation. But they both saw benefits. Dennett said his decision model could "give libertarians what they say they want." He was right, and it is surprising that more libertarians did not adopt Dennett's model and try to improve upon it, perhaps finding the proper role for quantum events, as the two-stage model has now done.

Mele's "agnostic autonomism" and "modest libertarianism" were designed to take the best parts of libertarian and compatibilist positions, and make them defensible whether determinism or indeterminism was "true."

Like Mele's models, the two-stage model is less "free" than extreme libertarian views, but more responsible. As Mele has said, in the second stage, the will is as adequately determined as any compatibilist could desire.

The two-stage model is also less "determined" than some extreme Compatibilist views, because it is not predetermined in the sense of a causal chain back to the

universe origin. But it is more creative than standard compatibilist views. It provides for adequate *determination* of the will by the agent's reasons, motives, feelings, and desires. But it also provides the limited indeterminism needed for the generation of new ideas that allow the agent to be the originator and author of her life.

David Hume reconciled freedom with determinism. We believe that the two-stage model reconciles free will with indeterminism.

Might compatibilists find this a satisfactory model for a more comprehensive compatibilism, one compatible both with adequate determinism *and* with indeterminism that is limited to the generation of alternative possibilities?

Of course the model is still *incompatible* with predeterminism, and it is distinct from the indeterminism after or centered at the moment of choice, including Kane's cases of "torn decisions."

The two-stage model is perhaps less "event-causal" and more "agent causal," because the agent has creative powers during the extended "moment of choice." These are the kind of powers sought by agent-causalist libertarians like Roderick Chisholm (Chisholm 1995), Richard Taylor (Taylor 1966), and Keith Lehrer (Lehrer 1966). These philosophers called for an absolute freedom, even from causes like reasons, motives, feelings, and desires. This shocked compatibilists at the time. Could such agent causalists be satisfied with the agent's ability to generate totally unconstrained new ideas right up to and including the "moment of choice," ideas that are not caused by anything prior to their generation?

Nothing in the events of the "fixed past" (and the laws of nature, as compatibilists like to say) up to the "moment of choice" *pre*determines the agent's decision. Because the first stage generates new alternative possibilities, the two-stage model lets the agent *choose otherwise* in exactly the same circumstances that obtained before the beginning of deliberation. Kane calls this the "Indeterminist Condition," he says "the agent should be able to act and act otherwise (choose different possible futures), given the same past circumstances and laws of nature" (Kane 2005).

This ability to do otherwise is often considered the most extreme requirement for libertarianism. The two-stage model now provides a credible explanation for this very important ability to do otherwise in exactly the same circumstances before the decision process began.

Discussions with Robert Kane at the Social Trends Institute's Experts Meeting in Barcelona and later have led to a convergence of views between Kane and the author. We both embrace indeterminism as an essential part of free will, the author in the first stage of my two-stage model, Kane in the late stage of a decision, where a choice between different options in a "torn decision" can involve indeterminism but without loss of responsibility.

16.12 Conclusion

Although the problem of free will is nearly twenty-three centuries old, it is time to acknowledge that today we have a plausible, practical, and scientific two-stage solution to the problem. About 125 years ago, William James said that we must

accept absolute chance as a part of that solution, comparing the role of chance explicitly to its role in evolution that Darwin had announced a quarter century earlier.

It has been a hundred years since William James's death, time for recognition of his great achievement, bravely proclaimed to an audience of Harvard Divinity School students in an age when chance was still considered atheistic and an affront to God's foreknowledge.

Seventy-five years ago, James's most important student, Dickinson Miller, writing under the pseudonym R.E. Hobart and just a few years after quantum indeterminacy was discovered, reminded us that *determination* by the will was also required (Hobart 1934). Unfortunately, Hobart's work was misread by many compatibilist philosophers as requiring *determinism*, not simply *determination*. Hobart explicitly denied *predeterminism*.

Fifty years ago, A.J. Ayer (Ayer 1954) and J.J.C. Smart (Smart 1961) perfected the standard logical argument against free will, that either determinism or indeterminism must be true, and that free will was impossible either way. If we are determined, we are not free. If we are undetermined, our will is random.

Just over a quarter century ago, Karl Popper, Henry Margenau, and Daniel Dennett discussed two-stage models for free will that connected random events to our decisions, but the general philosophical community remained determinist and compatibilist. This was despite Peter van Inwagen's *Consequence Argument* (van Inwagen, 1983), which denies free will if all our actions are traceable in a causal chain to events back long before we were born. And it was despite Robert Kane's book *Free Will and Values* (Kane 1984) which launched his campaign to find some intelligible way to make quantum indeterminacy the key to free will.

Now Martin Heisenberg has identified chance as generating alternative possibilities for action in the lowest animals. Evolution has no doubt conserved this ability to recruit chance, since it provides the significant biological advantage of creativity. Behavioral freedom in lower animals has evolved to become free will in higher animals and humans.

The two-stage model of first "free" and then "will" is simple, intuitive, and the common sense view of the layperson. Our thoughts *come to us* freely. Our actions *go from us* willfully.

We conclude that science is indeed compatible with our desire for human freedom.

References

Aristotle. (1933a). *Metaphysics, Book VI, 1027b12-14*. Cambridge, MA: Harvard Loeb Library.
Aristotle. (1933b). *Nichomachean ethics, Book III, v.6, 1113b19-22*. Cambridge, MA: Harvard Loeb Library.
Ayer, A. J. (1954). *Philosophical essays* (p. 275). New York: St. Martin's Press.
Baars, B. (1997). *In the theater of consciousness*. New York: Oxford University Press.
Balaguer, M. (2010). *Free will as an open scientific problem*. Cambridge, MA: MIT Press.

Chisholm, R. (1995). Agents, causes, and events: The problem of free will. In T. O'Connor (Ed.), *Agents, causes, and events: Essays on indeterminism and free will* (p. 95). Oxford: Oxford University Press.
Cicero, M. T. (1951). *De Natura Deorum*. Cambridge, MA: Harvard Loeb Library.
Compton, A. H. (1931). The uncertainty principle and free will. *Science, 74*, 1911.
Dennett, D. (1978). On Giving Libertarians What They Say They Want. In *Brainstorms* (p. 295). Montgomery, VT: Bradford Books.
Double, R. (1991). *The non-reality of free will*. New York: Oxford University Press.
Doyle, R. O. (2009). Free will: It's a normal biological property, not a gift or a mystery. *Nature, 459*, 1052.
Doyle, R. O. (2010). Jamesian free will. *William James Studies, 5*, 1–28.
Doyle, R. O. (2011). *Free will: The scandal in philosophy*. Cambridge, MA: I-PHI Press.
Eddington, A. S. (1958). *The nature of the physical world* (p. 294). Ann Arbor: University of Michigan Press.
Eddington, A. S. (1938). *The philosophy of physical science* (p. 182). New York: Macmillan.
Ekstrom, L. W. (2000). *Free will*. Boulder, CO: Westview Press.
Gazzaniga, M. (2011). *Who's in charge? Free will and the science of the brain*. New York: Harper-Collins.
Heisenberg, M. (2009). Is free will an illusion? *Nature, 459*, 164–165.
Hobart, R. E. (1934). Free will as requiring determination, and inconceivable without it. *Mind, Vol XLIII, 169*, 2.
Hume, D. (1975). Enquiries concerning Human Understanding, Section VI. In *Of liberty and necessity* (p. 95). Oxford: Clarendon Press.
James, W. (1880). Great Men, Great Thoughts, and the Environment. *Atlantic Monthly, 46*(276), 441–459.
James, W. (1956). The dilemma of determinism. In *The Will to Believe*. New York: Dover Press.
Kane, R. (1984). *Free will and values*. Albany, NY: SUNY Press.
Kane, R. (1999). Responsibility, luck, and chance. *The Journal of Philosophy, 96*(5), 225.
Kane, R. (2005). *A contemporary introduction to free will* (p. 265). New York: Oxford Press.
Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development, 5*, 183–191.
Layzer, D. (1975). The arrow of time. *Scientific American, 233*(6), 56–69.
Lehrer, K. (1966). An empirical disproof of determinism. In K. Lehrer (Ed.), *Freedom and determinism* (p. 175). New York: Random House.
Locke, J. (1959). *Essay concerning human understanding. Book II. Chapter XXI. Of power, Sections 14–21* (pp. 319–324). New York: Dover Press.
Long, A. A. (1986). *Hellenistic philosophy* (p. 149). Berkeley, CA: University of California Press.
Lucretius, T. (1982). *De Rerum Natura. Book 2, lines 251–262*. Cambridge, MA: Harvard Loeb Library.
Ludwig, G. (1953). Der Messprozess (The Process of Measurement). *Zeitschrift für Physik, 135*, 483.
Margenau, H. (1968). *Scientific indeterminism and human freedom*. Latrobe, PA: Archabbey Press.
Mayr, E. (1988). *Toward a new philosophy of biology* (p. 150). Cambridge, MA: Harvard Belknap Press.
Mele, A. (1995). *Autonomous agents* (pp. 212–213). New York: Oxford University Press.
Mele, A. (2006). *Free will and luck* (p. 9). Oxford: Oxford University Press.
Murphy, N., Ellis, G. F. R., & O'Connor, T. (2009). *Downward causation and the neurobiology of free will*. New York: Springer.
Nagel, T. (1979). Moral luck. In *Mortal Questions* (p. 24). Cambridge: Cambridge University Press.
Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.
Pereboom, D. (2001). *Living without free will*. Cambridge: Cambridge University Press.

Popper, K. (1977). *Natural selection and the emergence of mind*. Cambridge: Darwin College.
Smart, J. J. C. (1961). Free-will, praise and blame. *Mind, LXX*, 279.
Smilansky, S. (2000). *Free will and illusion*. Oxford: Clarendon.
Stebbings, L. S. (1958). *Philosophy and the physicists* (p. 185). New York: Dover Press.
Strawson, G. (1994). The impossibility of moral responsibility. *Philosophical Studies, 75*(1–2), 5–24.
Swinburne, R. (2011). Dualism and the determination of action. In R. Swinburne (Ed.), *Free will and modern science*. New York: Oxford University Press.
Taylor, R. (1966). *Action and purpose*. Englewood Cliffs, NJ: Prentice-Hall.
Van Inwagen, P. (2000). Free will remains a mystery. *Philosophical Perspectives, 14*, 14.
Wegner, D. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press.
Wiggins, D. (1973). Towards a reasonable libertarianism. In T. Honderich (Ed.), *Essays on freedom of action* (p. 33). London: Routledge & Kegan Paul.
Williams, B. (1981). *Moral luck* (p. 20). Cambridge: Cambridge University Press.

Chapter 17
Can a Traditional Libertarian or Incompatibilist Free Will Be Reconciled with Modern Science? Steps Toward a Positive Answer

Robert Kane

Abstract The landscape of free will debate was simpler in the 1960s when I first began dealing with the problem of free will. The unstated assumption was that if you had scientific leanings, you would naturally be a *compatibilist* about free will (believing it to be compatible with determinism). By contrast, if you defended a *libertarian* or *incompatibilist* free will, requiring indeterminism, you must inevitably reduce free will to mere chance or to the mystery of uncaused causes, immaterial minds, noumenal selves, or prime movers unmoved. The question I set for myself back then was how one might reconcile a traditional incompatibilist free will requiring indeterminism with modern science without reducing it to either chance or mystery. It has turned out that doing so required rethinking many facets of the traditional problem of free will from the ground up. I report on some results of this rethinking in this paper.

Keywords Free will • Incompatibilism • Libertarianism • Ultimate responsibility • Self-forming actions

17.1 Introduction

There has been a resurgence of interest in problems of the freedom of the will in the past half-century and many changes in dealing with them.

My own dealings with free will are coterminous with this resurgence and date back to the 1960s. The landscape of free will debate was simpler then. The unstated assumption was that if you had scientific leanings, you would naturally be a *compatibilist* about free will (believing it to be compatible with determinism). (That is, you would be a compatibilist if you were not a free will skeptic or hard

R. Kane (✉)
The University of Texas at Austin, Austin, TX, USA
e-mail: rkane@uts.cc.utexas.edu

A. Suarez and P. Adams (eds.), *Is Science Compatible with Free Will?: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience*,
DOI 10.1007/978-1-4614-5212-6_17,

determinist, denying that we had free will altogether.) If, by contrast, you were a *libertarian* about free will—that is to say, if you believed in a free will that is *incompatible* with determinism, as did many traditional thinkers—you must inevitably appeal to some obscure or mysterious forms of agency to make sense of it—to uncaused causes, immaterial minds, noumenal selves, nonevent agent causes, prime movers unmoved, or other examples of what P.F. Strawson called the "panicky metaphysics" of traditional defenders of libertarian free will (in his influential essay "Freedom and Resentment" (1962)).

I started thinking about free will shortly after Strawson's essay appeared, when my philosophical mentor at the time, Wilfrid Sellars, a well-known analytic philosopher of the period, challenged me to reconcile a traditional incompatibilist or libertarian free will with modern science. Sellars was a scientifically oriented thinker and he was a compatibilist about free will, like the vast majority of philosophers and scientists of that era. He did not believe a traditional free will that was incompatible with determinism could be accounted for without appealing to obscure or mysterious forms of agency of the kinds Strawson had dubbed "panicky metaphysics." Employing a well-known distinction he had introduced into the philosophical literature, Sellars granted that free will *in some sense* was an integral part of what he called "the manifest image" of humans and their world (our ordinary everyday view). But he did not believe a traditional free will that required incompatibilism could be reconciled with "the scientific image" of that world; and he challenged me to show otherwise.

I accepted the challenge at the time and remember thinking with the brashness and naivete of youth: "Give me three or four weeks and I'll wrap this up and be back with an answer (or at least by the end of the semester!)." Well, it is now more than forty years later and the effort is still ongoing. The reason the task was so much more difficult than I naively assumed was that, as I slowly came to realize, it required rethinking many facets of the traditional problem of free will from the ground up, breaking old molds of thought and substituting new ones. I report on some results of this rethinking in this paper. But first some historical background to the issues I will discuss.

17.2 Modernity and the Free Will Problem

> There is a disputation that will continue till mankind is raised from the dead, between the necessitarians and the partisans of free will.

These are the words of the thirteenth century Persian poet and Sufi thinker, Jalalu'ddin Rumi. The free will of which Rumi speaks is the traditional notion of freedom that many thinkers, Rumi included, have believed was in conflict with necessitarian or deterministic doctrines of all kinds—fatalistic, theological, physical, biological, psychological, and social. Yet that same traditional idea of free will of which he speaks—and which I believe to be incompatible with determinism—

has been under sustained attack in modernity as outdated, obscure and unintelligible and has been dismissed by many modern philosophers and scientists since the seventeenth century for its supposed lack fit with the modern images of the humans and the cosmos in the natural and human sciences. Nietzsche (1886) summed up a prevailing view in his inimitable prose when he said

> The desire for 'freedom of the will' in the superlative metaphysical sense. . .the desire to bear the ultimate responsibility for one's actions oneself. . .to be nothing less than a *causa sui*. . .is the best self-contradiction that has been conceived so far [by the mind of man]

I agree that this traditional idea of free will may appear utterly mysterious and obscure in a modern context unless we learn to think about it in new ways. Like many another issue of modernity, the question is whether something of the traditional idea of free will "in the superlative metaphysical sense" can be retrieved from the dissolving acids of modern science and secular learning or whether it will become, along with other aspects of our self-image, yet another victim of the "disenchantments" of modernity.

The doubts about this traditional notion of free will, expressed here by Nietzsche and shared by many modern thinkers, have deep roots. They are related to an ancient dilemma: If free will is not compatible with *determinism*, it does not seem to be compatible with *indeterminism* either. Determinism implies that, given the past and laws, there is only one possible future. Indeterminism implies the opposite: Same past and laws, different possible futures. On the face of it, indeterminism seems more congenial to the idea of an "open" future with branching pathways in decision-making—a "garden of forking paths," in the image of Jorge Luis Borges' well-known story of that name. But how is it possible, one might ask, that different actions or choices could arise voluntarily and intentionally from *exactly* the same past and (barring miraculous departures from the laws of nature) without occurring merely by luck or chance?

This question has had a hypnotic effect on those who think about free will. One imagines that if free choices were undetermined, then which occurs would be like spinning a wheel in one's mind or one must just pop out by chance or randomly. If, for example, a choice occurred by virtue of some undetermined quantum events in one's brain, it would seem a fluke or accident rather than a responsible choice. Such undetermined events occurring in our brains or bodies would not seem to enhance our freedom and control over our actions, but rather diminish our freedom and control, and hence our responsibility. Arguments of these kinds and many other now-familiar arguments have led through the centuries to familiar charges that undetermined choices, of the kind *incompatibilists* about free will demand, would be "arbitrary," "capricious," "random," "irrational," "uncontrolled," "inexplicable," "mere matters of luck or chance" and hence not really free and responsible choices at all.

It is little wonder that traditional libertarians about free will, who believed it was incompatible with determinism, have looked for some *deus ex machina* to solve the problem, while their opponents have cried magic or mystery. Indeterminism was required for free will, they argued, but it was not enough. It might provide causal

gaps in nature. But something else must fill those gaps. Some additional form of agency or causation was needed that went beyond causation in the natural order, whether deterministic or indeterministic. Thus, in response to modern science, there were numerous historical appeals in the modern era, from Descartes to Kant and beyond, to "extra factors" such as noumenal selves, immaterial minds, transempirical power centers, nonevent agent causes, uncaused causes, and the like, to account for a traditional libertarian or incompatibilist free will (Strawson's panicky metaphysics). I long ago became disenchanted with all such appeals.

17.3 Indeterminism and Agency

But where to go if one is to avoid such traditional strategies for explaining free will? I came to believe that one must take a new look at the issues from the ground up. First, let us be clear that it is an empirical and scientific question whether any indeterminism *is* there in nature in ways that are appropriate for free will—in the brain, for example. It may very well *not* be there; and in any case, no purely philosophical theory can settle the matter. As the Epicureans said centuries ago, if the atoms don't "swerve" in undetermined ways, and in the right places, there would be no room in nature for free will. I have long argued that the question of whether or not we have free will in appropriate senses has an empirical dimension and cannot be settled by a priori or philosophical reasoning alone. It cannot be settled, for example, as philosophers have sometimes assumed, by introspectively appealing to experiences of deliberating and choosing or by engaging in conceptual analysis of ordinary terms like "could" or "power." That is one reason why the free will issue has been so philosophically intractable. And it is why we philosophers need the aid of physicists, biologists, neuroscientists and other scientists, such as are gathered here at this conference, if we are to make progress on the issue.

Yet philosophical reasoning *is* relevant to many aspects of the free will problem. And our present question is the philosophical one that has boggled people's minds for centuries, from the time of the Epicureans onward: What could one *do* with indeterminism, assuming it *was* there in nature in the right places, to make sense of free will as something other than *mere* chance or randomness and without appealing to mystery? In other words, assume for the sake of argument that there was some indeterminism in nature in the right places, say, in the form of genuine randomness in the neural processing of rational agents, so that our behavioral modules involved, in the words of Martin Heisenberg, "an interplay between chance and lawfulness in the brain."[1] What could we do with this neural randomness to make sense of *human* free will, with its implications of rationality, autonomy, responsibility and moral agency, without reducing it to mere chance or mystery?

[1] See Chap. 7 in this book (Heisenberg 2013).

In the process of addressing this question, one would also be addressing another that is central to this conference: How might one get from the "randomly generated action" that Heisenberg postulates is characteristic of living things more generally, even in more primitive biological forms, to what the philosophers have traditionally called the "freedom of the will" in more complex rational, self-reflective beings like ourselves?

17.4 The Compatibility Question: Freedom, Responsibility and UR (Ultimate Responsibility)

To begin to address these questions, the first step is to ask what a traditional free will was supposed to involve and in particular why it was thought to be incompatible with determinism in the first place. We may begin to do this by reflecting on two more familiar notions to which free will is related—freedom and responsibility.

Nothing could be more important than freedom to the modern age. People clamor for it all over the world, often against authoritarian and violent resistance. And why do they want it? The simple, and not totally adequate, answer is that to be free is to be able to satisfy one's desires or do whatever one wants. In free societies, people can buy what they want, travel where they please, choose what to read, and so on. But these freedoms are what you might call *surface* freedoms. What we mean by *free will* runs deeper than these ordinary freedoms.

To see how, suppose we had maximal freedom to make choices of the above kinds to satisfy our desires, yet the choices we actually made were in fact manipulated by others, by the powers that be. In such a world we would have a great deal of everyday freedom to do whatever we wanted, yet our freedom of *will* would be severely limited. We would be free to *act* or to choose *what* we willed, but we would not have the ultimate power over what it is that we willed. Other persons would be pulling the strings, not by coercing or forcing us to do things against our wishes, but by manipulating us into having the wishes they wanted us to have. Indeed, this kind of "covert non-coercive (CNC) control," as I have called it in previous writings, is becoming the form of control of choice of the new millennium; and it is in some ways more sinister than coercive control. One sign of how important free will is to us is that people feel revulsion at such manipulation and feel demeaned by it when they find out it has been done to them. When subjected to it, they realize they were not their own persons; and having free will is about being your own person. We capture this in common parlance when we speak of acting "of our *own* free will."

Reflecting in this way on the differences between surface and deeper senses of *freedom* is one path to understanding the freedom of the will. Another is by reflecting on the notion of *responsibility*. Suppose a young man is on trial for an assault and robbery in which his victim was beaten to death. Let us say we attend his trial and listen to the evidence in the courtroom. At first, our thoughts of the young

man are filled with anger and resentment. What he did was horrible. But as we listen daily to how he came to have the mean character and perverse motives he did have—a sad story of parental neglect, child abuse, sexual abuse, bad role models—some of our resentment against the young man is shifted over to the parents and others who abused and mistreated him. We begin to feel angry with them as well as with him. Yet we aren't quite ready to shift all of the blame away from the young man himself. We wonder whether some residual responsibility may not belong to him. Our questions become: To what extent is *he* responsible for becoming the sort of person he now is? Was it *all* a question of bad parenting, societal neglect, social conditioning, and the like, or did he have any role to play in it?

These are crucial questions about free will and they are questions about what may be called the young man's *ultimate responsibility*. We know that parenting and society, genetic make-up and upbringing, have an influence on what we become and what we are. But were these influences entirely *determining* or did they "leave anything over" for us to be responsible for? That is what we want to know about the young man. The question of whether he is merely a victim of bad circumstances or has some residual responsibility for being what he is—the question, that is, of whether he became the person he is *of his own free will*—seems to depend on whether these other factors that were not under his control were or were not *entirely* determining.

Reflections such as these point to a basic condition that in my view has fueled intuitions that free will and determinism may be incompatible down through history. I call it the condition of *ultimate responsibility* or UR. The basic idea is this: to be ultimately responsible for an action, an agent must be responsible for anything that is a sufficient reason (condition, cause or motive) for the action's occurring.[2] If, for example, a choice issues from, and can be sufficiently explained by, an agent's character and motives (together with background conditions), then to be *ultimately* responsible for the choice, the agent must be at least in part responsible by virtue of choices or actions voluntarily performed in the past for having the character and motives he or she now has. Compare Aristotle's claim (1915: 1114a13-22) that if a man is responsible for wicked acts that flow from his character, he must at some time in the past have been responsible for forming the wicked character from which these acts flow.

This condition of ultimate responsibility or UR does not require that we could have done otherwise for *every* act done "of our own free wills." But it *does* require that we could have done otherwise with respect to *some* acts in our past life histories by which we formed our present characters. I call these *self-forming actions* or *SFAs*. Often we act from a will already formed, but it is "our own free will" by virtue of the fact that we formed it by other choices or actions in the past (self-forming actions or SFAs) for which we could have done otherwise. If this were not so, there is *nothing we could have ever done differently in our entire lifetimes to make ourselves different than we are*—a consequence, I believe, that is incompatible

[2] For a formal statement and defense of this condition, see Kane 1996, Chap. 3.

with our being (at least to some degree) ultimately responsible for what we are. So self-forming actions or SFAs are only a subset of those acts in life for which we are ultimately responsible and which are done "of our own free will." But if none of our acts were self-forming in this way, we would not be *ultimately* responsible for anything we did.

Focusing on UR tells us something else of importance about free will. It tells us why the free will issue is about the freedom *of the will* and not merely about the freedom of action. There has been a tendency in the modern era, beginning with Hobbes and Locke in the seventeenth century, to reduce the problem of free will to a problem of freedom of action. I have been arguing for some time that such a reduction oversimplifies the problem.[3] Free will is not just about free action. It is about *self-formation*, about the formation of our "wills" or how we got to be the kinds of persons we are, with the characters, motives and purposes we now have. Were *we* ultimately responsible to some degree for having the wills we do have, or can the sources of our wills be completely traced backwards to something over which we had no control—God or Fate, heredity and environment, nature or upbringing, society or culture, social conditioning or hidden controllers, and so on? Therein, I believe, lies the core of the traditional problem of "free will" of which Rumi and others have spoken.

Finally, and no less importantly, focusing on UR tells us why free will has been thought to be incompatible with determinism. If agents must be responsible to some degree for anything (such as their prior formed character and motives) that is a sufficient cause or motive for their actions, an impossible infinite regress of past actions would be required unless some actions in an agent's life history ("self-forming actions") did not have either sufficient causes or motives and hence were undetermined.

17.5 Indeterminism and Responsibility

But this new route to incompatibility of free will and determinism raises a host of further questions about free will, including how actions lacking both sufficient causes and motives could themselves be free and responsible actions, and how, if at all, such actions could exist in the natural order where we humans live and have our being. My own first efforts at dealing with this problem in the 1970s, as Bob Doyle notes in his contribution to this volume, was to formulate a two-stage model very much like the one he develops in his paper. The idea was that in the process of deliberation, various thoughts, memories, images, etc. would come to mind in undetermined and unpredictable ways (the first stage) and these undetermined events would then influence the outcome of the deliberation, namely which choice was the more rational or preferable one to make (the second stage). (Such a

[3] Kane 1985, 1989, 1994, 1996, 1999, 2002, 2005, 2009.

two-stage view was first suggested by William James, as Doyle points out in his paper, and has since been developed by others besides myself, including Daniel Dennett [1978] and Alfred Mele [1995], as well as by Doyle himself [2011].)

I thought from the beginning that this must be a part of the solution to the free will problem. But I also believed that it could not be the complete solution. The reason was that this two-stage model did not fully capture the deep kind of responsibility (i.e., ultimate responsibility) that genuine free will requires. And this was owing to the fact that which choice turns out to be the more rational or preferable on this two-stage model would depend on which undetermined thoughts, etc. have occurred earlier in the deliberation and which have not. Yet which undetermined thoughts occurred earlier in the deliberation, and which did not, would be a matter of chance and not something over which the agent had control. By contrast, ultimate responsibility requires that for at least some choices in our life histories the indeterminism must occur at the moment of choice itself, and not merely earlier in deliberation. (I do not deny that it is deeply puzzling how this could be without reducing the choice *itself* to mere chance. But this is the challenge that must be faced, I believe, if a full account of free will is to be given.) As a result, while I made the two-stage model part of my own theory in my first book on free will (1985), it was only a part of the theory and I also tried to go beyond it.

I am even more convinced today through the work of Martin Heisenberg (reported in his contribution to this volume) as well as these others just mentioned, including Mele and Doyle, that not only is the two-stage model an important part of any adequate theory of free will, but that it is also an important, indeed a crucial, step in the evolution of human free will. The ability to randomize in lower organisms affords them flexibility and creativity as it does for humans. But I believe, as I did in the 1970s, that a number of other steps are needed to get from this first crucial evolutionary step to the full evolution of free will in human beings, and that the two-stage model must be folded into a larger picture.

So I turn now to this larger picture. The first step is to note, as indicated earlier, that indeterminism does not have to be involved in all acts done "of our own free wills" for which we are ultimately responsible. Not all such acts have to be undetermined, but only those by which we made ourselves into the kinds of persons we are, namely "self-forming actions" or SFAs.

Now I believe these undetermined self-forming actions occur at those difficult times of life when we are torn between competing visions of what we should do or become. Perhaps we are torn between doing the moral thing or acting from ambition, or between powerful present desires and long-term goals, or we are faced with a difficult task for which we have aversions. In all such cases, we are faced with competing motivations and have to make an effort to overcome temptation to do something else we also strongly want. There is tension and uncertainty in our minds about what to do at such times, I suggest, that may be reflected in appropriate regions of our brains by further far-from-equilibrium behavior—in short, a kind of "stirring up of chaos" in the brain that makes it sensitive to micro-indeterminacies at the neuronal level. The uncertainty and inner tension we feel at such soul-searching moments of self-formation would then be reflected in the indeterminacy of our

neural processes themselves. What is experienced internally as uncertainty would correspond physically to the opening of a window of opportunity that would temporarily screen off *complete* determination by influences of the past. (By contrast, when we act from settled motives and character, the uncertainty or indeterminacy would be muted or damped.)

If we were to decide under such conditions of uncertainty, the outcome would not be determined because of the preceding indeterminacy—and yet it could be willed (and hence rational and voluntary) either way owing to the fact that in such self-formation, the agents' prior wills are divided by conflicting motives. Consider a businesswoman who faces such a conflict. She is on her way to an important meeting when she observes an assault taking place in an alley. An inner struggle ensues between her conscience, to stop and call for help, and her career ambitions which tell her she cannot miss this meeting. She has to make an effort of will to overcome the temptation to go on. If she overcomes this temptation, it will be the result of her effort, but if she fails, it will be because she did not *allow* her effort to succeed. And this is due to the fact that, while she willed to overcome temptation, she also willed to fail, for quite different and incommensurable reasons. When we, like the woman, decide in such circumstances, and the indeterminate efforts we are making become determinate choices, we would *make* one set of competing reasons or motives prevail over the others then and there *by deciding*.

Now add a further piece to the puzzle. Just as indeterminism need not undermine rationality and voluntariness, so indeterminism in and of itself need not undermine control and responsibility. Suppose you are trying to think through a difficult problem and there is some indeterminacy in your neural processes complicating the task—a kind of chaotic background. It would be like trying to concentrate and solve a problem, say a mathematical problem, with background noise or distraction. Whether you are going to succeed in solving the problem is uncertain and undetermined because of the distracting neural noise. Yet, if you concentrate and solve the problem nonetheless, there is reason to say you did it and are responsible for it even though it was undetermined whether you would succeed. The indeterministic noise would have been an obstacle that you overcame by your effort.

There are numerous examples supporting this point, first suggested by J.L. Austin, Elizabeth Anscombe and others in 1960s, where indeterminism functions as an obstacle to success without precluding responsibility. Consider an assassin who is trying to shoot a prime minister, but might miss because of some undetermined events in his nervous system that may lead to a wavering of his arm. If the assassin does succeed in hitting his target, despite the indeterminism, can he be held responsible? The answer is clearly yes because he intentionally and voluntarily succeeded in doing what he was *trying* to do—kill the prime minister. Yet his action, killing the prime minister, was undetermined. Or, here is another example: a husband, while arguing with his wife, in a fit of rage swings his arm down on her favorite glass table top intending to break it. Again, we suppose that some indeterminism in his outgoing neural pathways makes the momentum of his arm indeterminate so that it is genuinely undetermined whether the table will break right up to the moment when it is struck. Whether the husband breaks the table or not is undetermined and yet he is clearly

responsible if he does break it. It would be a poor excuse for him to say to his wife: "chance did it, not me." Even though there was a chance he wouldn't break it, chance didn't do it, *he* did.

Now these examples—of the mathematical problem, the assassin and the husband—are not all we want, since they do not amount to genuine exercises of (self-forming) free will in SFAs, like the businesswoman's, where the will is divided between conflicting motives. The assassin's will is not divided between conflicting motives as is the woman's. He wants to kill the prime minister, but does not also want to fail. (If he fails therefore, it will be *merely* by chance.) Yet these examples of the assassin, the husband and the like, do provide some clues. To go further, we have to add further thoughts.

17.6 Parallel Processing

Imagine in cases of inner conflict characteristic of SFAs, like the businesswoman's, that the indeterministic noise which is providing an obstacle to her overcoming temptation is not coming from an external source, but is coming from her own will, since she also deeply desires to do the opposite. Imagine that two crossing (recurrent) neural networks are involved, each influencing the other, and representing her conflicting motivations. (Recurrent networks, as we know, are complex networks of interconnected neurons circulating impulses in feedback loops that are generally thought to be involved in higher-level cognitive processing.[4]) The input of one of these neural networks consists in the woman's reasons for acting morally and stopping to help the victim; the input of the other, her ambitious motives for going on to her meeting.

The two networks are connected so that the indeterministic noise which is an obstacle to her making one of the choices is coming from her desire to make the other, and vice versa—the indeterminism thus arising from a tension-creating conflict in the will, as I said. In these circumstances, when either of the pathways reaches an activation threshold (which amounts to choice), it will be like your solving the mathematical problem by overcoming the background noise produced by the other. And just as when you solved the mathematical problem by overcoming the distracting noise, one can say you did it and are responsible for it, so one can say this as well in the present case, I would argue, *whichever one is chosen*. The pathway through which the woman succeeds in reaching a choice threshold will have overcome the obstacle in the form of indeterministic noise generated by the other.

Note that, under such conditions, the choices either way will not be "inadvertent," "accidental," "capricious," or "merely random" (as critics of incompatibilist freedom say), because they will be *willed* by the agents either way when they are

[4] Accessible introductions to the role of such neural networks in cognitive processing include Churchland 1996 and Spitzer 1999. For more advanced discussion, see Churchland and Sejnowski 1992.

made, and done for *reasons* either way—reasons that the agents then and there *endorse*. But these are the conditions usually required to say something is done "on purpose," rather than accidentally, capriciously or merely by chance. Moreover, these conditions taken together, as I have argued elsewhere, rule out each of the reasons we have for saying that agents act, but do not have *control* over their actions (compulsion, coercion, constraint, inadvertence, accident, control by others, etc.).[5]

Indeed, in these cases, agents have what I call "*plural voluntary control*" over the options in the following sense: They are able to bring about *whichever* of the options they will, *when* they will to do so, for the *reasons* they will to do so, on *purpose* rather than accidentally or by mistake, without being coerced or compelled in doing so or willing to do so, or otherwise controlled in doing or willing to do so by any other agents or mechanisms. I show in my 1996 book (Chaps. 8–10) that each of these conditions can be satisfied for SFAs as conceived above, even though the SFAs are undetermined. The conditions can be summed up by saying, as we sometimes do, that the agents can choose either way *at will*.

Note also that this account of self-forming choices amounts to a kind of "doubling" of the mathematical problem. It is as if an agent faced with such a choice is *trying* or making an effort to solve *two* cognitive problems at once, or to complete two competing (deliberative) tasks at once—in our example, to make a moral choice and to make a conflicting self-interested choice (corresponding to the two competing neural networks involved). Each task is being thwarted by the indeterminism that is due to the presence of the other, so it might fail. But if it succeeds, then the agents can be held responsible because, as in the case of solving the mathematical problem, they will have succeeded in doing what they were willingly trying to do. Recall the assassin and the husband. Owing to indeterminacies in their neural pathways, the assassin might miss his target or the husband fails to break the table. But if they *succeed*, despite the probability of failure, they are responsible, because they will have succeeded in doing what they were trying to do.

And so it is, I suggest, with self-forming choices or SFAs, except that in the case of self-forming choices, *whichever way the agents choose* they will have succeeded in doing what they were trying to do because they were simultaneously trying to make both choices, and one is going to succeed. Their failure to do one thing is not a *mere* failure, but a voluntary succeeding in doing the other.

Does it make sense to talk about agents trying to do two competing things at once in this way, or to solve two cognitive problems at once? Well, much current scientific evidence points to the fact that the brain is a parallel processor; it simultaneously processes different kinds of information relevant to tasks such as

[5] We have to make further assumptions about the case to rule out some of these conditions. For example, we have to assume, no one is holding a gun to the woman's head forcing her to go back, or that she is not paralyzed, etc. But the point is that the satisfaction of these further conditions is consistent with the case of the woman as we have imagined it. If these other conditions are satisfied, as they can be, and the business woman's case is in other respects as I have described it, We have an SFA. I offer the complete argument for this in Kane 1996, Chap. 8, among other works listed in Note 2.

perception or recognition through different neural pathways. Such a capacity, I believe, is essential to the exercise of free will. In cases of self-formation (SFAs), agents are simultaneously trying to resolve plural and competing cognitive tasks. They are, as we say, of two minds. Yet they are not two separate persons. They are not dissociated from either task. The businesswoman who wants to go back to help the victim is the same ambitious woman who wants to go to her meeting. She is torn inside by different visions of who she is and what she wants to be, as we all are from time to time. But this is the kind of complexity needed for genuine self-formation and free will. And when she succeeds in doing one of the things she is trying to do, she will endorse that as *her* resolution of the conflict in her will, voluntarily and intentionally, not by accident or mistake.[6]

Note also that these reflections give us the beginning of an answer to the further question asked earlier of how might one get from the "randomly generated action" that Heisenberg postulates is characteristic of many living things, to what the philosophers have traditionally called the "freedom of the will" in more complex beings like ourselves. Such randomly generated action in living things would provide an evolutionary template for the development of free will. But what one would have to add to it are the rational and reflective capacities to imagine different possible ways of acting and different visions of who one might be—that is, to imagine, as I put it earlier, creatures like ourselves who could be from time to time (figuratively) "of two (or more) minds, without being two separate persons." Being such creatures would not merely require intelligent behavior and acting in pursuit of values. It would in addition require capacities for higher-order valuation (evaluation of the values one pursues) and hence higher-order reflection about who one is and what one wants to be. Such capacities were assigned by the ancients to "Reason" and to beings possessing it.

17.7 Responsibility, Luck, and Chance

Now you may find all this interesting and yet still find it hard to shake the intuition that if choices are undetermined, they *must* happen merely by chance—and so must be "random," "capricious," "uncontrolled," "irrational," and all the other things

[6] Another related objection that is commonly made at this point is that it is irrational to make efforts to do incompatible things. I concede that in most ordinary situations it is. But I argue that there can be special circumstances in the deliberative lives of rational agents in which it is not irrational to make competing efforts: These include circumstances in which (i) we are deliberating between competing options; (ii) we intend to choose one or the other, but cannot choose both; (iii) we have powerful motives for wanting to choose each of the options for different and incommensurable reasons; (iv) there is a consequent resistance in our will to either choice, so that (v) if either choice is to have a chance of being made, effort will have to be made to overcome the temptation to make the other choice; and most importantly, (vi) we want to give each choice a fighting chance of being made because the motives for each choice are important to us; and we would taking them lightly if we did not make an effort in their behalf. These conditions are the conditions of SFAs.

usually charged. Such intuitions are deeply ingrained and they give rise to a host of questions and objections that naturally arise and have been made about the view just presented.

The first step in exorcising deeply ingrained intuitions about indeterminism is to question the intuitive connection in most people's minds between "indeterminism's being involved in something" and "its happening merely as a matter of chance or luck." "Chance" and "luck" are terms of ordinary language that carry the connotation of "its being out of my control." So using them already begs certain questions, whereas "indeterminism" is a technical term that merely precludes *deterministic* causation, though not causation altogether. Indeterminism is consistent with non-deterministic or probabilistic causation, where the outcome is not inevitable. It is therefore a mistake (alas, one of the oldest and most common in debates about free will) to assume that "undetermined" means "uncaused." (Libertarian freedom was often characterized in the past, wrongly I believe, as "contra-causal" freedom.)

Here is another source of misunderstanding. Since the outcome of the businesswoman's effort (the choice) is undetermined up to the last minute, one may have the image of her first making an effort to overcome the temptation to go on to her meeting and then at the last instant "chance takes over" and decides the issue for her. But this is misleading. One cannot separate the indeterminism and the effort of will, so that *first* the effort occurs *followed* by chance or luck (or vice versa). Rather the effort *is* indeterminate and the indeterminism is a property of the effort, not something separate that occurs after or before the effort. The fact that her effort has this property of being indeterminate does not make it any less the woman's *effort*. The complex recurrent neural network that realizes the effort in the brain is circulating impulses in feedback loops and there is some indeterminacy in these circulating impulses. But the whole process is her effort of will and it persists right up to the moment when the choice is made. There is no point at which the effort stops and chance "takes over." She chooses as a result of the effort, even though she might have failed. Similarly, the husband breaks the table as a result of his effort, even though he might have failed because of the indeterminacy. (That is why his excuse, "chance broke the table, not me" is so lame.)

Just as expressions like "she chose *by* chance" can mislead in such contexts, so can expressions like "she got lucky." Recall that, with the assassin and husband, one might say "they got lucky" in killing the prime minister and breaking the table because their actions were undetermined. *Yet they were responsible*. So ask yourself this question: why does the inference "he got lucky, so he was *not* responsible?" fail in the cases of the husband and the assassin where it does fail? The first part of an answer to this question has to do with the point just made that "luck," like "chance," has question-begging implications in ordinary language that are not necessarily implications of "indeterminism." The core meaning of "he got lucky" in the assassin and husband cases, which *is* implied by indeterminism, I suggest, is that "he succeeded *despite the probability or chance of failure*"; and this core meaning does not imply lack of responsibility, *if he succeeds*.

The second reason why the inference "he got lucky, so he was not responsible" fails for the assassin and the husband is that *what* they succeeded in doing was what

they were *trying* and wanting to do all along (kill the minister and break the table respectively). The third reason is that *when* they succeeded, their reaction was not "oh dear, that was a mistake, an accident—something that *happened* to me, not something I *did*." Rather they *endorsed* the outcomes as something they wanted all along, and did so knowingly and purposefully, not by mistake or accident.

But these conditions are satisfied in the businesswoman's case as well, *either way* she chooses. If she succeeds in choosing to return to help the victim (or in choosing to go on to her meeting) (i) she will have "succeeded *despite the probability or chance of failure*," (ii) she will have succeeded in doing what she was *trying* and *wanting* to do all along (she wanted both outcomes very much, but for different reasons, and was trying to make those reasons prevail in both cases), and (iii) when she succeeded (in choosing to return to help) her reaction was not "oh dear, that was a mistake, an accident—something that happened to me, not something I did." Rather she *endorsed* the outcome as her resolution of the conflict in her will. And if she had chosen to go on to her meeting she would have endorsed that outcome, recognizing it as her resolution of the conflict in her will.

Another objection often made to the preceding view is that we are not introspectively aware of making dual efforts and performing multiple cognitive tasks in such choice situations. But I am not claiming that agents are conscious of making dual efforts. What they are introspectively conscious of is that they are trying to decide about which of two options to choose and that either choice is a difficult one because there are resistant motives pulling them in different directions that will have to be overcome, whichever choice is made. In such introspective conditions, I am theorizing that what is actually going on underneath is a kind of parallel distributed processing in the brain that involves separate efforts or endeavorings to resolve competing cognitive tasks.

This is an example of a point made earlier that introspective evidence cannot give us the whole story about free will. Stay on the surface and things *are* likely to appear obscure or mysterious. What is needed is a *theory* about what might be going on behind the scenes when we exercise free will, not merely a description of what we immediately experience; and in this regard new scientific ideas can be a help rather than a hindrance to making sense of the subject. If parallel distributed processing takes place on the *input* side of the cognitive ledger (in perception), then why not consider that it also takes place on the *output* side (in practical reasoning and choice)? That is what we should suppose, I am suggesting, if we are to make sense of incompatibilist free will.

It has also been objected that indeterminism would undermine the notion of *agency* itself by turning choices and actions into mere chance events. As noted earlier, that worry sends us scurrying around looking for extra factors, other than prior events or happenings, to tip the balance to one choice or the other. But there is an alternative way to think about the way that indeterminism might be involved in free choice, a way that avoids these familiar libertarian stratagems and requires a transformation of perspective.

The idea is not to think of the indeterminism involved in free choice as a cause *acting on its own*, but as an *ingredient* in a larger goal-directed or teleological

process or activity in which the indeterminism functions as a *hindrance* or *obstacle* to the attainment of the goal. Such is the role I have suggested for indeterminism in the efforts preceding undetermined SFAs.

We tend to reason that if an outcome (breaking a table *or* making a choice) depends on whether certain neurons fire or not (in the arm *or* in the brain), then the agent must be able to *make* those neurons fire or not, if the agent is to be responsible for the outcome. In other words, we think we have to crawl down to the place where the indeterminism originates (in the individual neurons) and *make* them go one way or the other. We think we have to become originators at the micro-level and tip the balance that chance leaves untipped, if we (and not chance) are to be responsible for the outcome. And we realize, of course, that we can't do that. But we don't have to. It's the wrong place to look. We don't have to micromanage our individual neurons one by one to perform purposive actions and we do not have such micro-control over our neurons *even when we perform ordinary actions* such as swinging an arm down on a table.

What we need when we perform purposive activities, mental *or* physical, is macro-control of processes involving many neurons—processes that may succeed in achieving their goals despite the interfering effects of some indeterminacies in the processing. We do not micro-manage our actions by controlling individual neurons or muscles and it would be counterproductive to try. But that does not prevent us from macro-managing our purposive activities (whether they be mental activities, such as practical reasoning, or physical activities, such as arm-swingings) and being responsible for those activities when they succeed.

17.8 Responsibility and Control

But does not the presence of indeterminism or chance at least *diminish* the control persons have over their choices or actions? Is it not the case that the assassin's control over whether the prime minister is killed (his ability to realize his purposes or what he is trying to do) is lessened by the undetermined impulses in his arm—and so also for the husband and his breaking the table? The answer is yes. But the further surprising point worth noting is that *diminished control* in such circumstances does not entail *diminished responsibility* when the agents succeed in doing what they are trying to do.

Ask yourself this question: Is the assassin less guilty of killing the prime minister, if he did not have complete control over whether he would succeed because of the indeterminism in his neural processes? Suppose there were three assassins, each of whom killed a prime minister. Suppose one of them had a fifty percent chance of succeeding because of the indeterministic wavering of his arm. Another had an eighty percent chance, and the third, a young stud, nearly a hundred percent chance. Is one of these assassins less guilty than the other, *if they all succeed*? Should we say that one assassin deserves a hundred years in jail, the other eighty years and the third fifty years? Absurd. They are all equally guilty if they succeed. The diminished

control in the assassins who had an eighty percent or a fifty percent chance does not translate into diminished responsibility when they succeed. Diminished control in such circumstances does not entail diminished responsibility.

There is an important further lesson here about free will in general. We should concede that indeterminism, wherever it occurs, *does* diminish control over what we are trying to do and *is* a hindrance or obstacle to the realization of our purposes. But recall that in the case of the businesswoman (and SFAs generally), the indeterminism that is diminishing her control over one thing she is trying to do (the moral act of helping the victim) *is coming from her own will*—from her desire and effort to do the opposite (go to her business meeting). And the indeterminism that is diminishing her control over the other thing she is trying to do (act selfishly and go to her meeting) is coming from her desire and effort to do the opposite (to be a moral person and act on moral reasons). In each case, the indeterminism *is* functioning as a hindrance or obstacle to her realizing one of her purposes—a hindrance or obstacle in the form of resistance within her will which has to be overcome by effort.

If there were no such hindrance—if there were no resistance in her will—she would indeed in a sense have a more "complete control" over one of her options. There would be no competing motives standing in the way of her choosing it and therefore no interfering indeterminism. But then also, she would not be free to rationally and voluntarily choose the other purpose because she would have no good competing reasons to do so. Thus, by *being* a hindrance to the realization of some of our purposes, indeterminism paradoxically opens up the genuine possibility of pursuing other purposes—of choosing or doing *otherwise* in accordance with, rather than against, our wills (voluntarily) and reasons (rationally). To be genuinely self-forming agents (creators of ourselves)—to have free will—there must at times in life be obstacles and hindrances in our wills of this sort that we must overcome. Self-formation is not a gift, but a struggle.[7]

Of interest here is Kant's image, which I have used before, of the bird that is upset by the resistance of the air and the wind to its flight and so imagines that it could fly better if there were no air at all to resist it. But, of course, as Kant points out, the bird would not fly better if there were no air. It would cease to fly at all. So it is with indeterminism in relation to free will. It provides resistance to our choices, but a resistance that is necessary if we are to be capable of genuine self-formation.

17.9 Liberum Arbitrium

I conclude with one final objection. Even if one granted that persons, such as the businesswoman, could make genuine self-forming choices that were undetermined, isn't there something to the charge that such choices would be *arbitrary*? A residual

[7] If one were to take a religious perspective, this fact might be related to the problem of evil. Compare Evodius's question to St. Augustine, in Augustine's classic work on free will (Augustine 1964), of why God gave us free will since it brings so much conflict, struggle and suffering into the world.

arbitrariness seems to remain in all self-forming choices since the agents cannot in principle have sufficient or conclusive *prior* reasons for making one option and one set of reasons prevail over the other.

There is some truth to this objection as well, but again I think it is a truth that tells us something important about free will. It tells us that every undetermined self-forming free choice is the initiation of what might be called a value experiment whose justification lies in the future and is not fully explained by past reasons. In making such a choice we say, in effect, "Let's try this. It is not required by my past, but it is consistent with my past and is one branching pathway my life can now meaningfully take. Whether it is the right choice, only time will tell. Meanwhile, I am willing to take responsibility for it one way or the other."

The term "arbitrary," as I have often noted, comes from the Latin *arbitrium*, which means "judgment"—as in *liberum arbitrium voluntatis*, "free judgment of the will" (the medieval philosophers' designation for free will). Imagine a writer in the middle of a novel. The novel's heroine faces a crisis and the writer has not yet developed her character in sufficient detail to say exactly how she will act. The author makes a "judgment" about this that is not determined by the heroine's already formed past which does not give unique direction. In this sense, the judgment (*arbitrium*) of how she will react is "arbitrary," but not entirely so. It had input from the heroine's fictional past and in turn gave input to her projected future. In a similar way, agents who exercise free will are both authors of and characters in their own stories all at once. By virtue of "self-forming" judgments of the will (*arbitria voluntatis*) (SFAs), they are "arbiters" of their own lives, "making themselves" out of past that, if they are truly free, does not limit their future pathways to one.

Suppose we were to say to such persons: "But look, you didn't have sufficient or *conclusive* prior reasons for choosing as you did since you also had viable reasons for choosing the other way." They might reply. "True enough. But I did have *good* reasons for choosing as I did, which I'm willing to stand by *and take responsibility for*. If these reasons were not sufficient or conclusive reasons, that's because, like the heroine of the novel, I was not a fully formed person before I chose (and still am not, for that matter). Like the author of the novel, I am in the process of writing an unfinished story and forming an unfinished character who, in my case, is myself."

References

Aristotle. (1915). Nichomachean ethics. In W. D. Ross (Ed.), *The works of Aristotle* (Vol. 9). London: Oxford University Press.

Augustine. (1964). *On the free choice of the will*. Indianapolis: Bobbs-Merrill.

Churchland, P. M. (1996). *The engine of reason, the seat of the soul*. Cambridge, MA: MIT Press.

Churchland, P. S., & Sejnowski, T. J. (1992). *The computational brain*. Cambridge, MA: MIT Press.

Dennett, D. (1978). *Brainstorms*. Cambridge, MA: MIT Press.

Doyle, B. (2011). *Free will: The scandal of philosophy*. Cambridge, MA: I-PHI Press.

Heisenberg, M. (2013). The origin of freedom in animal behaviour. In A. Suarez & P. Adams (Eds.), *Is Science compatible with Free Will?* New York: Springer. Chapter 7.
Kane, R. (1985). *Free will and values*. Albany, NY: State University of New York Press.
Kane, R. (1989). Two kinds of incompatibilism. *Philosophy and Phenomenological Research, 31*, 219–254.
Kane, R. (1994). Free will: The elusive ideal. *Philosophical Studies, 75*, 25–60.
Kane, R. (1996). *The significance of free will*. Oxford: Oxford University Press.
Kane, R. (1999). Responsibility, luck and chance: reflections on free will and indeterminism. *The Journal of Philosophy, 96*, 217–240.
Kane, R. (2002). Some neglected pathways in the free will labyrinth. In R. Kane (Ed.), *The oxford handbook of free will* (pp. 406–437). Oxford: Oxford University Press.
Kane, R. (2005). *A contemporary introduction to free will*. Oxford: Oxford University Press.
Kane, R. (2009). Free will and the dialectic of selfhood. *Ideas Y Valores, 58*, 25–44.
Mele, A. (1995). *Autonomous agents*. New York: Oxford University Press.
Nietzsche, F. (1886). Jenseits von Gut und Böse, I, 21. Leipzig: Naumann; *Beyond Good and Evil. I*, 21 Cambridge UK: Cambridge University Press, 2002. http://www.gutenberg.org/files/4363/4363-h/4363-h.htm#2HCH0001, Cited 28 October 2012.
Spitzer, M. (1999). *The mind within the net*. Cambridge, MA: MIT Press.
Strawson, P. F. (1962). Freedom and Resentment. *Proceedings of the British Academy, 48*, 1–25.

Chapter 18
Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience

Peter Adams and Antoine Suarez

Abstract This chapter attempts to show that the different observations and arguments presented in this book, while coming from quite different disciplines, are related and complement each other. They support the conclusion that quantum physics and neuroscience are compatible with free will and consciousness. What is more, these seem to be becoming basic irreducible principles (axioms) of today's science: Consciousness and free will undoubtedly exist, and they must be a fundamental ingredient of any sound explanation of the world. Knowledge, and in particular science, cannot be thought of as separated from the domain of human rights and responsibilities. The ultimate reason for choosing free will may be the profound desire of ensuring personal identity and authorship, and so making it possible to claim personal rights.

The synthesis proposed in this chapter is the sole responsibility of the authors. Nevertheless, we think that it may help explain why the interdisciplinary communications between the experts at the STI Meeting that has given rise to this book succeeded beyond expectations.

Keywords Indeterminism • Nonlocality at detection • Wake–sleep cycle • Free will • Limited consciousness • Self-formation • Rights • Responsibility

P. Adams (✉)
Thomas More Institute, 18b Netherhall Gardens, London NW3 5TH, UK
e-mail: peter.adams@thomasmoreinstitute.org.uk

A. Suarez
Center for Quantum Philosophy, The Institute for Interdisciplinary Studies, Berninastr. 85, 8057 Zurich, Switzerland

Social Trends Institute/Bioethics, Barcelona, Spain
e-mail: suarez@leman.ch

A. Suarez and P. Adams (eds.), *Is Science Compatible with Free Will?: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience*,
DOI 10.1007/978-1-4614-5212-6_18,

18.1 Introduction

From the arguments advanced in the different contributions in this book one can derive the following main results:

1. Free will and consciousness are taken more and more as axioms or basic principles of science. Consciousness and free will are actually *irreducible*, that is, they cannot be explained by other things but rather are necessary ingredients in any scientific explanation. Science is based on observation, and the results of experiments are the greatest scientific authority. However, observation cannot be defined other than through relation to an observer. This means that scientific knowledge is based on the way humans become aware of data coming from the external world.
2. Regarding free will and consciousness, "quantum nonmaterial agency" coming from outside space-time may be a much more relevant concept than "quantum indeterminism." Acknowledging that phenomena can be influenced by nonmaterial agency, and that even the visible accessible world emerges from invisible, inaccessible powers, science remains open to principles and concepts like freedom, personal identity, creativity, responsibility and religious faith. This is actually a historic result that reveals an emerging social trend: Science seems to be getting back the freedom and mind that went missing two centuries ago.
3. Human consciousness and free will are limited. The misinterpretation of Libet's experiments as denying free will is based on the prejudice that human free will is always accompanied by full consciousness. Libet's experiments actually support the idea that voluntary actions in humans can exhibit many degrees of consciousness, going from unconscious voluntary actions to highly conscious ones: A conscious action may be preceded by an unconscious preparation of the brain; or one can now with full consciousness agree to perform an action later, and then actually realize it without consciously deciding when precisely to do it; or one can even have the illusion of having chosen to do something but, as Tononi (2013) states, "the existence of illusions of will, just as that of visual illusions, does not imply that free will in general is illusory, or that visual experience is unreal."
4. Free will and consciousness are eminently accessible through first-person knowledge. I conclude that some other human performs an action with a certain intention because he performs the same kind of movements I perform to reach a particular goal.
5. Free will and consciousness are involved in many physical phenomena from the very elementary level of quantum devices in the lab to the level of animal behavior. However, rights and responsibility can be considered distinctive human features. This means that the sense of rights and responsibility is crucial for making our knowledge of the world consistent.

In the following we attempt to show how these results relate to the different contributions.

18.2 Indeterminism, Quantum Nonlocality, and Free Will

Quantum indeterminism means essentially that the result of quantum experiments cannot be explained by a chain of causes reaching from the present moment back to the origin of the Universe. In this sense it is clear that without indeterminism one could not have freedom in the world. One should not forget that free will is not something separated from the body: Free decisions happen in the brain, and the brain is part of nature and functions according to the principles of physics, for example the conservation of energy. Had nature been deterministic, the neuronal dynamics of my brain could be fully explained by causes in the past, and free will would be impossible.

This does not mean that indeterminism explains free will, but only establishes that indeterminism is a necessary condition to have a world where freedom is possible. Gilles Brassard expresses this point well in his contribution: "I also acknowledge the difficulty of deriving free will from probabilities, randomness and nondeterminism. Nevertheless, I am inhabited by an unshakable belief that free will, if it exists, cannot have another origin, with apologies to the compatibilists" (Brassard and Raymond-Robichaud 2013, Sect. 4.8). And similarly Giulio Tononi says: "The requirement for indeterminism implies that, even though a choice may feel free [...], if we knew for certain that a choice is completely preordained due to absolute determinism, we would conclude that the feeling of responsibility is an illusion. The choice would indeed be autonomous—ours and nobody else's; [...] but though fully and irreducibly, consciously ours, it would also be inevitable" (Tononi 2013, Sect. 11.4).

On the other hand indeterminism is not a sufficient condition for free will: "Many concede that some degree of indeterminism is essentially guaranteed, not only due to quantum phenomena but simply to the unpredictability of the environment. [...] However, at least since David Hume, it has been argued that this kind of indeterminism does nothing to assuage the feeling that responsibility is ultimately illusory: to the extent that a choice is determined, ultimate responsibility remains an illusion, and to the extent that it is indeterminate or random, it becomes merely arbitrary" (Tononi 2013).

Free will is surely more than indeterminism: it is a power capable of influencing and controlling to some extent visible phenomena, and in particular the neuronal dynamics of our brains. And this is why quantum nonlocality matters so much for free will, as Nicolas Gisin stresses (Gisin 2013). Indeed nonlocality shows us how randomness already at the level of "inanimate devices" does not mean a lack of control, but a low level of control coming from outside space-time. Quantum physics is telling us that the following principle is basic for science:

Principle Q: "not all that is important for physical phenomena is contained in space-time."

Antonio Acín in Chap. 2 impressively states that quantum physics takes as a main axiom that observers have free will. Then the observation of nonlocal correlations implies the randomness of the outcomes, i.e., outcomes which cannot

be explained by any narrative in space-time and therefore are unpredictable in principle. On the one hand, randomness cannot be proven "from scratch": one cannot guarantee the presence of randomness without resorting to some initial seed of randomness, that is, without invoking free will. On the other, the randomness seen in quantum phenomena is not simply a consequence of the initial assumed randomness: "new non-previously existing randomness is generated by the quantum setup" (Acín 2013). So here more than explaining free will by randomness one does exactly the opposite, in fact. Randomness appears to be a particular case of free will.

In the same line of thinking the quite recent experiments demonstrating nonlocality at detection presented in Chap. 5 are of great relevance: They put nonlocality in relationship with the conservation of energy. These experiments show that the most basic principle ruling the material world, *the conservation of energy*, could not hold without nonlocality, i.e., the material visible world emerges from nonmaterial invisible things (Suarez 2013).

All these results from quantum physics seem to fit in well with the theory of consciousness proposed by the psychiatrist and neuroscientist Giulio Tononi: A conscious choice is both "maximally and irreducibly causal" and "also necessarily under-determined and thus unpredictable." However, "in this view, indeterminism is not to be thought of as a sprinkle of randomness that instills some arbitrariness into a preordained cascade of mechanisms, decreasing their causative powers." This means in the end that "conscious causality" is not something that can be explained by any information which is stored in space-time but has to be conceived of as nonmaterial agency (coming from outside space-time) and responsible for the fact that "a complex at a macro-scale in space or time (groups of neurons, hundreds of milliseconds)" appears as "a maximal integrated information," chosen among other alternative possibilities at a certain moment. In this view indeterminism provides a ground state of the neuronal dynamics that can be controlled by free will (acting from outside space-time) to become an integrated state of information (Tononi 2013).

18.3 "Many Worlds"

The "parallel lives" theory presented by Gilles Brassard and Paul Raymond-Robichaud in Chap. 4 brings into focus an important point. These authors state that the experimental violation of Bell's inequalities rules out theories assuming "local hidden variables" but cannot be considered a proof of nonlocality (Brassard and Raymond-Robichaud 2013). Thereby Brassard and Raymond-Robichaud confirm John Bell's feeling that the "many worlds picture" is a further development of de Broglie's "empty wave," and has something distinctive to say regarding the quantum correlations (Bell 1987, p. 194).

Adrian Kent (2010) has put at the head of his meticulous criticism of "many worlds" the following quotation from P.K. Feyerabend: "...so crowded with...empty sophistication that it is extremely difficult to perceive the simple errors at its basis. It is like fighting the hydra—cut off one ugly head, and eight formulations take its place." Indeed Kent's criticism itself is a self-fulfilling prophecy: Vincent Duhamel and Paul Raymond-Robichaud have replied to Kent showing that the objection he raises against "many worlds" points rather to "a fundamental limitation of probabilities and statistics" and holds also for theories assuming a single-world (Duhamel and Raymond-Robichaud 2011).

If one accepts the "empty wave" it is impossible "to fight the hydra." Accepting "empty waves" means accepting entities that exist and propagate within space-time but that are not directly accessible to general observation: "empty waves" interact with the environment only in a very selective and specific way—actually an "empty wave" does not interact with any particle (and for this reason cannot be detected), but only with "its particle." But this is the very assumption on which "parallel lives" is based: space–time may be subdivided into many compartments ("bubbles"), which can interact with the environment only in a selective and predetermined way. If one accepts that at beam-splitter BS a particle P splits into a particle P' (leaving by output-port BS1) and an empty wave W' (leaving by output-port BS2), one can as well assume that the split produces additionally the alternative outcome, that is, a second empty wave W^* (leaving by output-port BS1) and a particle P^* (leaving by output-port BS2): The particle & wave pair ($P'\&W'$) is within the bubble Alice', and the pair ($P^*\&W^*$) within the parallel bubble Alice*, and both bubbles cannot interact with each other; however the bubble Alice'—containing the outcome ($P'\&W'$)—may interact with (depending on the experiment) either the bubble Bob'—containing the outcome ($P'\&W'$)—or the bubble Bob*—containing the outcome ($P^*\&W^*$) (Brassard and Raymond-Robichaud, 2013, p. 57). Hence, if one accepts "empty waves" it is impossible to oppose the "parallel lives" version of "many worlds."

The analysis by Brassard and Raymond-Robichaud shows the importance of the following principle:

> *Principle A:* All that is in space-time is accessible to observation (except in the case of space-like separation).

If you are going to be consistent, you are only capable of opposing the "many worlds" view if you accept *Principle A*.

In summary, the analysis by these authors leads to the conclusion that if one assumes decision of outcome at the beam-splitter, and thereby one accepts "empty waves," one cannot actually incorporate free will in the theory and prove nonlocality. Thus they indirectly stress that nonlocality at detection is the genuine form of nonlocality and in this sense is more basic than Bell's nonlocality.

However, no matter how intellectually gratifying a world picture may be in which all things evolve unitarily according to "many worlds," we prefer with Nicolas Gisin (2012, 2013) and Sandu Popescu (2012) to have a science that is not completely unitary if it allows me to claim my freedom and defend my rights.

18.4 Two-Stage Model, Indeterminism, and Nonmaterial Free Will

The contributions by Robert Kane and Bob Doyle are inspired by the wish to overcome Hume's objection that to the extent that a choice is indeterminate or random, it becomes merely arbitrary and ultimate responsibility remains an illusion (Tononi 2013; Merali 2013). So the crucial question is: "how might one reconcile a traditional incompatibilist free will requiring indeterminism with modern science without reducing it to either chance or mystery" (Kane 2013).

For Doyle indeterminism marks a first stage of free will, which permits it to happen. But for free will actually to happen, determination is required in a second stage. Kane describes the model in these terms: "The idea was that in the process of deliberation, various thoughts, memories, images, etc., would come to mind in undetermined and unpredictable ways (the first stage) and these undetermined events would then influence the outcome of the deliberation, namely, which choice was the more rational or preferable one to make (the second stage)" (Kane 2013).

Both Doyle and Kane embrace indeterminism as an essential part of free will: Doyle in the first stage of his Two-Stage Model, Kane in the late stage of a decision, where a choice between different options in a "torn decision" can involve indeterminism but without loss of responsibility (Doyle 2013).

However, in perfect agreement with the conclusion of the quantum physicists, Doyle and Kane assume that free-willed actions require something more than indeterminism.

Referring to cases in which we are faced with competing motivations and have to make an effort to overcome the temptation to do something else we also strongly want, Kane states: "There is tension and uncertainty in our minds about what to do at such times, I suggest, that may be reflected in appropriate regions of our brains by further far-from-equilibrium behavior—in short, a kind of 'stirring up of chaos' in the brain that makes it sensitive to micro-indeterminacies at the neuronal level. The uncertainty and inner tension we feel at such soul-searching moments of self-formation would then be reflected in the indeterminacy of our neural processes themselves. What is experienced internally as uncertainty would correspond physically to the opening of a window of opportunity that would temporarily screen off complete determination by influences of the past" (Kane 2013).

Kane apparently states that the indetermination at the level of free will induces indetermination at the level of brain dynamics and thereby the self cannot be considered only a product of the past.

We think that here philosophy and quantum physics meet: As we have seen in Sect. 18.2 quantum randomness does not mean "complete lack of order," but rather reveals influences coming from outside space-time. Quantum randomness itself can be considered a particular case of free will. Thus, it is the same will that can operate in different ways, going from unconscious operations to fully conscious ones. Only when the will acts fully consciously do we have "the deep kind of responsibility that genuine free will requires." It is the kind of responsibility that induces "self-formation" of the character (Kane 2013; Merali 2013).

In any case assuming free will as a nonmaterial principle acting from outside space-time should today no longer be considered "reducing free will to mystery," but rather a basic principle of science. Paraphrasing Lewis (1947) one could say that quantum physics offers the possibility of nonphysical agents acting on physical reality, and in particular opens the door to free will and consciousness.

18.5 Consciousness and the Wake–Sleep Cycle

Any meaningful message can be characterized by a finite number of bits exhibiting a particular statistical distribution. Suppose for instance that a digitized sentence expressing some wish of mine contains 40% of "1" and 60% of "0." It is natural to assume that when I utter such a sentence I arrange the physiological parameters of my brain in a way that is suitable for producing the required distribution 40% of "1" and 60% of "0." This capability of self-influencing the physiological parameters of the brain in order to produce a desired distribution of outcomes is what characterizes the state of wakefulness.

The relevance of self-control to conscious voluntary behavior is stated by Giulio Tononi as follows: "The requirement for *self-control* implies that, to be free, one must be able to influence one's choices. That is, merely registering some state of affairs but not being able to influence its outcome does not allow for freedom." According to his IIT [Integrated Information Theory] a system that could categorize its own past states without any ability to affect its own future states would not form a complex and could not generate voluntary conscious outcomes (Tononi 2013).

In Chap. 5 it has been argued that conscious and free-willed control of outcomes from outside space-time is in principle possible to produce meaningful pieces of information. Nevertheless this could in principle be done independently of the particular settings (path-length difference) of the quantum device (interferometer). Indeed the distribution quantum mechanics predicts for the prepared quantum state is supposed to hold for "a large number" of outcomes but the theory doesn't establish how many outcomes are needed to have "a large number" of them. In this sense conscious and free-willed behavior would not be tied to quantum physics, or in other words, there is no need of quantum physics to explain consciousness and free will beyond the fact that the neuronal dynamics is not completely predetermined by any information stored in space-time.

However, the capability of human beings for conscious self-control is fundamentally limited, not by any quantum mechanical principle but by the need to sleep. So human consciousness is well characterized by stating that "it is what vanishes every night when we fall into dreamless sleep and reappears when we wake up or when we dream" (Tononi 2008).

Sleep means temporary lack of self-control. When we dream, especially during REM (rapid eye movement) sleep, we may experience highly emotional narratives (Koch 2009). Nonetheless we do not adapt purposefully the outcomes of our brains to the real context around us, and without muscle atonia (temporary central

paralysis) during REM we would unconsciously act out our dreams and sometimes exhibit uncontrolled violent behavior (Koch 2009). When we sleep the brain produces random outcomes (Hobson 2005), which can be compared—partially—to the outputs a quantum interferometer produces in the lab.

To be awake means to be able to control the distribution of the outcomes and adapt them, in a consistent manner, to the real environment. Accordingly one can speculate that the wake–sleep cycle requires a quantum world: Sleep is the axiom, and the quantum the theorem, rather than the other way around.

Sara Gonzalez, Stephen Perrig, and Rolando Grave de Peralta stress in Chap. 10 that consciousness cannot be reduced to the activation of a few brain areas but one should rather think of consciousness "as a highly dynamic and emergent property of complex systems" (Gonzalez et al. 2013).

However, the wake–sleep cycle indicates that the capability for conscious self-control requires the activation of some neuronal system in the brain, which (on the basis of available experience) may possibly be located in the Ascending Reticular Activating System (ARAS). And pushing the comparison with the interferometer model of the brain further one could say that the neurophysiology of the ARAS accounts for the connection between the detectors and the switch allowing the experimenter to change the optical path length (see Figure 1a in Chap. 5). A short sequence of outcomes can act like a key capable of changing the switch to a position which permits a distribution of outcomes tuned to produce the desired meaning. A brain functions like a device expanding meaning: A small seed of meaning is expanded into a much longer string of meaning. This expansion of meaning is the counterpart of the expansion of randomness presented by Antonio Acín in Chap. 2.

At the same time, sense data is able to influence the physiological parameters of the brain as well. So when I perceive some behavior outside of me, the areas which become activated in my brain are related to those which become activated when I perform the same actions myself, in accordance with what observations of the mirror system show (Fogassi and Rizzolatti 2013).

The hypothesis of self-regulation of the brain parameters accounts for the relevant fact that we experience ourselves as free beings capable of *mental effort* to produce purposeful behavior (like speaking and acting). But it is clear that for the time being we have no idea about how to sketch out a precise relationship between mental effort and outcome distributions imposed by physiological parameters. Neither can we explain how the repetition of conscious and deliberate decisions drives the growth of connections between brain cells, and leads to the creation of personal skills and habits within processes of learning, character building, moral and civil education, etc. Another open question is why sleep proceeds in several cycles of alternating non-REM and REM stages (Hobson 2005; Tononi 2008).

In summary, the speculation we have made in this section is not an attempt to explain sleep by quantum mechanics, but rather the opposite, to explain quantum physics by sleep. Producing a device capable of limiting consciousness (namely, the brain of the human observer) may be a key step in the evolution of the Universe, and in this sense sleep may become a fundamental principle of science in a world ruled by quantum physics.

18.6 Voluntary Inhibition

The quantum philosophical view presented up to now does not invalidate any law of physics, and additionally overcomes the dualistic view of the soul "as something divorced from the tangible grey matter" (Suarez 2011).

When someone plays the piano his/her spiritual powers are not responsible for generating the energy necessary to trigger the movement of his/her fingers but rather for the order in which the fingers move. We could represent this order in digital form by a very large sequence of bits (1,0,0,1,1,0,1,0,0,...). This bit-string could represent for instance the digital transcription of the motor activity necessary to play a Beethoven Sonata. This order may be more or less deliberate. I often surprise myself by moving my fingers and hands as if I were playing the piano, without any deliberation. However, the sequence of movements contains also information corresponding to a sequence of bits (0,1,0,0,1,0,1,1,...) obviously different of that corresponding to the Beethoven Sonata. What I say is that in both cases the order of the bits in the string originates from my spiritual powers.

Similarly, when I whistle a song I am controlling more or less consciously my breathing. Again the particular sequence of movements (not the energy needed to trigger and sustain them) comes from my spiritual powers. When I am sleeping, the sequence of my respiratory movements even if it has no specific meaning, comes from my spiritual powers as well although with a very low level of consciousness. I regulate the movement of my lips through my free will when I move them consciously and when I move them without being aware of doing this.

In this sense random spontaneous movements of a human body are nothing other than a particular expression of human free will: they reveal unconscious free will, they are unconscious voluntary movements. As Martin Heisenberg stresses: "Behavioral freedom" is possible without consciousness, and there are situations in which I am conscious but not free. "For my actions to be free I do not have to be conscious of them" (Heisenberg 2013), even if I have to be conscious of them to be responsible for them. The relationship between spontaneous movements and Thomas Aquinas' concept of imperfect voluntary movements has been discussed (Suarez 2011).

The view that the will is somehow involved in the random generation of spontaneous movements fits well with the clinical criteria adopted for defining death: if breathing, eye and leg movements directly reveal a nonmaterial spiritual power, one cannot deduce the death of a person as long as his body shows such movements. One could even think that certain neural diseases constrain the free will so much that there are movements commanded by the will that are somewhat "involuntary" (purposeless).

The intrinsic relationship between free will and spontaneous movements is suggested in the contribution of Flavio Keller and Jana M. Iverson in Chap. 8, according to which the voluntary inhibition of spontaneous behavior (internal reflexes) in humans "appears to be a prerequisite for the emergence of free will" (Keller and Iverson 2013). One could say for instance that during rapid eye movement (REM) sleep the eyes move spontaneously in a random way without any

inhibition, while in the state of wakefulness conscious control of the brain's output inhibits undesired behavioral patterns and makes meaningful behavior possible.

More generally Keller and Iverson conclude: "This ability to inhibit behavioral patterns that are inconsistent with our long-term goals gives us the ability to sustain a specific course in life, despite countless stimuli and adverse events that might otherwise let us deviate from our intended course. Such behavior is typically human and is compatible with the existence of free will."

18.7 Spontaneous Movements and Mirror Neurons

Actually I experience directly only my own consciousness and free will: These are accessed through first-person knowledge. How would I conclude that the human being in front of me also shares these valuable capabilities? This is another crucial question related to the neurophysiological discovery of mirror neurons presented by Leonardo Fogassi and Giacomo Rizzolatti in Chap. 9. The mirror system unifies action production and action observation, thus allowing an understanding of the actions of others from the inside (first-person knowledge) (Fogassi and Rizzolatti 2013).

I understand the actions of the others I perceive, on the basis of my own actions. The neurophysiological correlate of this philosophical supposition is the mirror neurons system: "Each time an individual observes another individual performing an action, the set of neurons that encode that action is activated in the observer's own cortical motor system" (Rizzolatti and Sinigaglia 2010).

When I look at a colleague speaking during lunch, I conclude that the human form sitting in front of me is a person because: (a) it has the same specific form (or shape) as me, and (b) this form exhibits movements like the movements I make when expressing my thoughts, emotions, and claims for rights.

"Movements like the movements I make to express my feelings and claim for my rights" is what we call *spontaneous movements*. It is irrelevant whether such movements are conscious or not, because the distinction between conscious and unconscious is not always sharp, mainly regarding the movements performed by others (Suarez 2011).

Accordingly the specificity of the body and its spontaneous movements are decisive in grounding interpersonal communication, and can be considered the observable basis of rights.

18.8 Libet's Experiments

Several contributors to this book discuss the Benjamin Libet experiments (Libet 2002) from different perspectives. For the sake of clarity we summarize the experimental result as follows: The subject flexes his wrist at time 0; the arousal of the readiness potential in the brain is measured at time -550 ms (550 ms before time 0); the subject declares that he has taken the decision to flex at time -200 ms (200 ms before time 0).

Javier Bernácer and José Manuel Giménez Amaya in Chap. 12 advance the relevant observation "that an activation of the premotor or motor cortex is not always followed by a movement." An intriguing example of this comes from mirror neurons, which are activated when a subject performs an action, and also when that action is simply observed (Fogassi and Rizzolatti 2013). The example shows that activation of premotor and motor cortices cannot be considered a sufficient cause to make an actual movement, even in a case where the subject can be considered to perceive the action in a state of awareness (Bernácer and Giménez Amaya 2013). With even more reason, in the case of Libet's subjects who are supposed to lack awareness of the action they perform, the readiness potential at time -550 ms should not be considered a sufficient cause for the flexing of the wrist the subjects perform.

On similar lines Al Mele in Chap. 13 argues that Libet's setup exhibits a loophole due to the fact that in the absence of a muscle burst, there is no record of the brain's activity. So on the occasions when there were no flexes one cannot discount that the readiness potential occurred in Libet's subjects as usual starting at -550 ms and lasting about 300 ms until the subject *consciously* decided not to flex. These subjects provide a proof that the readiness potential alone does not cause the flexing, but is only an unconscious causal preparation of the "conscious proximal decision" to flex occurring at time -200 ms. Therefore, the available data does not contradict the hypothesis that the flexing is caused by a "conscious proximal intention," and this specific causal process "includes no unconscious proximal decisions or intentions to flex" (Mele 2013).

Jean Staune in Chap. 14 stresses that even Libet himself admits the possibility that free will can be exercised by vetoing the urge to flex after the person declares he is aware of this urge. And he concludes: "Something fundamental happens 0.2 s before the act. This is the moment where the "I" or the "self" have a chance to *stop or to continue* the processes which have been started without it." It is noteworthy that sharing this conclusion Staune disagrees with Libet to some extent, since Libet in fact establishes a window for the veto lasting from -150 ms to -50 ms (Mele 2013), and agrees rather with Mele's analysis in Chap. 13. According to Staune the idea that the "self" has a chance to *consciously* stop or to continue a process which started unconsciously, corresponds to our everyday experience: We make a lot of movements without really being aware of them. It is the case with hand movements during lively discussions. We can however "take control of our bodies at any moment by crossing our arms and keeping our hands still." Accordingly free will is no illusion, although it is more limited than one often assumes (Staune 2013).

This argument may be strengthened by the "voluntary inhibition" Keller and Iverson (2013) describe in Chap. 8. Indeed one could think that the fact that the flexing happens depends of the subject's "conscious proximal intention" not to voluntarily inhibit the flexing at time −200 ms.

Nonetheless, even if it would turn out that in Libet's experiment the cause of the decision is an unconscious activity of the brain and the subject's feeling of having decided freely is an illusion, this would not be a final argument against free will, as Giulio Tononi suggests: "the existence of illusions of will, just as that of visual illusions, does not imply that free will in general is illusory, or that visual experience is unreal" (Tononi 2013).

Rather than questioning free will, Benjamin Libet's experiments address the issue of whether we should be considered responsible for an action we have consciously decided to perform at a later occasion, but then when the occasion arrives we perform the action in an unconscious voluntary way.

On the one hand one can state with Tononi: "a choice is the freer, the more it is conscious" (Tononi 2013).

But on the other hand one can say with Kane that "ultimate responsibility [. . .] does not require that we could have done otherwise for every act done 'of our own free wills.' But it does require that we could have done otherwise with respect to some acts in our past life histories by which we formed our present characters." In the case of Libet's experiments a self-forming action is performed at the moment the subject freely decides to participate in the experiment. Once he is involved in the experiment he acts "from a will already formed," but it is his own free will by virtue of the fact that he formed it by other choices or actions in the past [. . .] for which he could have done otherwise (Kane 2013).

For this reason we think the economist Luis Cabral is right when in Chap. 15 he questions the idea that, if we are able to measure brain activity well enough, then economic behavior will be predictable, and believes there is an irreducible degree of uncertainty which results from each individual's free will: "My main point is that the statistical regularity of aggregate economic behavior is compatible with irreducible uncertainty and unpredictability of individual behavior; and that the latter results from individual free will. In many ways, the point I am making about economics can also be made about other human and social sciences. The reason for my particular focus is that, among the social sciences, economics is the field that comes closest to the idea of a deterministic model of the sort offered by classical mechanics" (Cabral 2013).

In summary, voluntary movements are not always conscious and deliberate. Libet's experiments refute neither free will nor personal responsibility, but rather demonstrate that human consciousness and purposeful free will are limited.

18.9 Are Humans the only Free Agents in the Universe?

As we have stressed in Sect. 18.2 if one accepts the free will of the experimenter, then quantum experiments demonstrate the effects in which control of quantum randomness happens from outside space-time. In this sense such experiments can be considered an experimental proof of free will *on the part of nature* (assuming free will *on the part of the experimenter*) in accordance with Anton Zeilinger's view of the "two freedoms" (Zeilinger 2005).

If we accept that there is a place for free will in human brains, then quantum experiments imply that there is free will in nature outside human brains. This is the content of the so-called free will theorem Zeeya Merali presents in Chap. 6 (Merali 2013). And Heisenberg, Martin argues in Chap. 7 that animals also exhibit behavioral freedom (Heisenberg 2009, 2013). It is clear that not only humans but also

animals show spontaneous movements. If one assumes that human spontaneous movements reveal nonmaterial agency coming from outside the space-time, it seems consistent to accept that the spontaneous movements of nonhuman animals reveal nonmaterial agency as well (Suarez 2011).

Additionally, if one claims that "consciousness is integrated information" (Tononi 2008, 2013), then one is led to conclude that certain animal behavior and even quantum phenomena involve consciousness.

But if humans are not the only free and conscious agents in the universe, where does the freedom and consciousness in the universe outside "human experimenters" come from?

Regarding this question three positions seem possible (Suarez 2011):

1. The elementary particles all over the universe and nonhuman animals share free will and consciousness like humans do. This position appears more or less explicitly in certain formulations of "The free will theorem": "If we (humans) have free will then so do elementary particles" (Conway and Kochen 2006, 2009; Merali 2013).
2. The behavior of elementary particles and nonhuman animals is guided by divine agency.
3. The behavior of elementary particles and nonhuman animals requires agency coming from outside space-time that is neither of divine origin nor like that of a human soul. In particular, elementary particles and nonhuman animals cannot be considered *bodies* of the agents controlling them from outside space-time like the human body is supposed to be the body of a human soul.

In the closing talk of the Seminar on nonlocality at the occasion of his 60th birthday Nicolas Gisin stated: "There must be some register tracking the status of 'who is entangled with whom' (similar to a register of who is married with whom)." And he asked whether we have to accept "angels who keep track of the quantum register" (Gisin 2012).

Following Lewis (1947) again one could say that quantum physics offers the possibility of nonphysical agents acting on physical reality. In fact all major civilizations refer to "angels," and important philosophical traditions following Plato, Aristotle, and Thomas Aquinas claim that "in this visible world nothing takes place without the agency of the invisible intellects" (Aquinas, STh I, 110.1). Hence the issue deserves to be discussed in more depth and looks like a fascinating philosophical challenge for the coming years (Suarez 2007).

18.10 Rights and Responsibility as Distinctive Human Features

But if consciousness and free will pervade the whole universe, what features can be considered distinctive for humans? (This looks like a second major philosophical challenge science is proposing to us.)

In several chapters of this volume there are elements that may be useful for tackling this challenge.

For instance, Martin Heisenberg in Chap. 7 stresses the importance of "shared intentionality" for communication and cooperation in human societies: "Among humans the issue of freedom occurs predominantly in the social context." In animal societies the needs of the group are imposed upon the individual. In species such as Drosophila freedom is not as important an issue as in humans because "the quality of fly behavior is not compromised as much by the fly's social interactions as is that of human behavior."

Similarly, Antoine Suarez in Chap. 5 highlights concepts like "personal identity," "authorship," and "personal rights." For defining the person and her rights both "individuality" and "specificity" seem to play a key role. As stated in Sect. 18.7, I conclude that a human body is a person because it has the same specific form (or shape) as mine, and it exhibits movements like the movements I make when expressing my thoughts, emotions, and claims for rights (spontaneous movements).

The best way I can ensure that I am respected by others is to assume that spontaneous movements in a body of the human species reveal personal agency, and making this assumption the basis of my assigning rights to others. A human body that performs movements like the movements I make to express my feelings and rights-claims is a person I have to respect. Otherwise, I cannot rationally claim that he should presume to respect me.

Civil and penal law, for instance, actually assume that the behavior of a human body is the observable basis for deciding about its rights and responsibility. Rights originate from the will to grant to bodies of the human species that they respect each other. It is primarily because one wishes to explain human bodiliness and organize human society on the basis of rights, that one derives concepts like animation and life, and one applies them subsequently and somewhat by analogy to animals, which are often characterized as organisms capable of spontaneous and voluntary motion. In this context it is worth asking whether it is still possible to grant a rational foundation for rights, if one totally disposes of the human bodily architecture as an observable basis for defining rights.

According to other contributors "responsibility" is the characteristic of human behavior. So for instance Giulio Tononi in Chap. 11 Sect. 11.14 states that "a choice can only be free if it cannot be ascribed to anything less than oneself—we are the only entity that can be said to be responsible for the choice. That is, when asking who is responsible for the choice, the answer should be 'me,' meaning all the circuits underlying my present conscious experience, and nothing less than that."

Bob Doyle in Chap. 16 Sect. 16.10 considers that chance can only generate random (unpredictable) alternative possibilities for action or thought. By contrast: "The choice or selection of one action must be adequately determined, so that we can take responsibility. And once we choose, the connection between mind/brain and motor control must be adequately determined to see that 'our will be done.'"

As we have seen in Sect. 18.8, Robert Kane in Chap. 17 Sect. 17.4 introduces the "condition of ultimate responsibility": "to be ultimately responsible for an action,

an agent must be responsible for anything that is a sufficient reason (condition, cause, or motive) for the action's occurring. If, for example, a choice issues from, and can be sufficiently explained by, an agent's character and motives (together with background conditions), then to be ultimately responsible for the choice, the agent must be at least in part responsible by virtue of choices or actions voluntarily performed in the past for having the character and motives he or she now has" (Kane 2013).

Independently of whether particles and animals share free will or they are guided from outside space-time according to statistical rules by invisible intellects, it does not make sense to pretend that these agents (the particles, animals, or whatever intellects behind them) share "responsibility" for the natural effects (sometimes real catastrophes) they may bring about. Rights and responsibility seem to make sense only within the frame of the human species, that is, they are defined within the species.

Robert Kane points out another feature that may be at the core of human freedom: "Now I believe these undetermined self-forming actions occur at those difficult times of life when we are torn between competing visions of what we should do or become. Perhaps we are torn between doing the moral thing or acting from ambition, or between powerful present desires and long term goals, or we are faced with a difficult task for which we have aversions. In all such cases, we are faced with competing motivations and have to make an effort to overcome temptation to do something else we also strongly want."

Along the same lines Tononi refers to the tension between "a general concept of right or wrong" I may have, and many other concepts I also have "of what is advantageous to me."

If rights are an intraspecific human feature and require an observable basis, it seems natural to accept belonging to the human species as the principle defining the person and the foundation of personal rights. Occasionally I may have the power to do things that are advantageous to me but would harm others. But if I harm others and thereby I act against the foundation of rights, I will not be able to claim that I be respected myself.

We think it is the tension between the requirements of some fundamental principle to which we reasonably adhere, and other competing particular interests or presumed advantages that make the essence of human freedom. All during my life "I am in the process of writing an unfinished story and forming an unfinished character who, in my case, is myself" (Kane 2013). Human freedom means in the end that I have the possibility of forming myself as a person or contradicting myself.

18.11 Conclusion

The observations and arguments presented in this book clearly support the conclusion that quantum physics and neuroscience are perfectly compatible with free will and consciousness. What is more, these are taken today as basic irreducible principles

(axioms) of science. Paraphrasing Giulio Tononi one could state that consciousness and free will undoubtedly exist, and must be a fundamental ingredient of any sound explanation of the world—as fundamental as energy and space-time (see Tononi 2008, 2013).

Metaphysics is based on observation, and today's science provides experiments that lead to challenging philosophical questions beyond the scientific realm and may inspire metaphysical reflection.

Knowledge and in particular science cannot be thought of as separated from the domain of human rights and responsibility. The ultimate reason for choosing free will may be the profound desire of ensuring personal identity and authorship, and so making it possible to claim personal rights. Freedom of will must be presupposed as a quality of all rational beings (Kant 1785). No matter how intellectually gratifying a world picture may be in which all things evolve unitarily according to "many worlds," we prefer to have a science that is not completely unitary if it allows me to claim my freedom and defend my rights.

References

Acín, A. (2013). True quantum randomness. In A. Suarez & P. Adams (Eds.), *Is science compatible with free will?* New York: Springer. Chapter 2.

Aquinas, T. (STh I 110.1). Summa theologica I. Q. 110.1, on the contrary. http://www.newadvent.org/summa/1110.htm#article1. Cited 27 June 2012.

Bell, J. S. (1987). *Speakable and unspeakable in quantum mechanics*. Cambridge: Cambridge University Press.

Bernácer, J., & Giménez Amaya, J. M. (2013). On habit learning in neuroscience and free will. In A. Suarez & P. Adams (Eds.), *Is Science compatible with Free Will?* New York: Springer. Chapter 12.

Brassard, G., & Raymond-Robichaud, P. (2013). Can free will emerge from determinism in quantum theory? In A. Suarez & P. Adams (Eds.), *Is Science compatible with Free Will?* New York: Springer. Chapter 4.

Cabral, L. (2013). Are the laws of economics compatible with free will? In A. Suarez & P. Adams (Eds.), *Is Science compatible with Free Will?* New York: Springer. Chapter 15.

Conway, J., & Kochen, S. (2006). The free will theorem. *Foundations of Physics, 36*, 1441. quant-ph/0604079.

Conway, J., & Kochen, S. (2009). The strong free will theorem. *Notices of the American Mathematical Society, 56*, 226232. arXiv: 0807.3286.

Doyle, B. (2013). The two-stage solution to the problem of free will. In A. Suarez & P. Adams (Eds.), *Is Science compatible with Free Will?* New York: Springer. Chapter 16.

Duhamel, V., Raymond-Robichaud, P. (2011). Guildenstern and Rosencrantz in quantumland. A reply to Adrian Kent. http://arxiv.org/abs/1111.2563.

Fogassi, L., & Rizzolatti, G. (2013). The mirror mechanism as a neurophysiological basis for action and intention understanding. In A. Suarez & P. Adams (Eds.), *Is Science compatible with Free Will?* New York: Springer. Chapter 9.

Gisin, N. (2013). Are there quantum effects coming from outside space-time? Nonlocality, free will and "no many-worlds". In A. Suarez & P. Adams (Eds.), *Is Science compatible with Free Will?* New York: Springer. Chapter 3.

Gisin, N. (2012). Closing talk at the Seminar: "The Greatest Inspiration Surely Is Nonlocality". Vall d'Illiez/Switzerland, June 1 (cited according to private communication of Mi 13.06.2012 14:00).

Gonzalez, S. L., Perrig, S., & Grave de Peralta, R. (2013). On the quest for consciousness in vegetative state patients through electrical neuroimaging. In A. Suarez & P. Adams (Eds.), *Is Science compatible with Free Will?* New York: Springer. Chapter 10.

Heisenberg, M. (2009). Is free will an Illusion? *Nature, 459*, 164–165.

Heisenberg, M. (2013). The origin of freedom in animal behaviour. In A. Suarez & P. Adams (Eds.), *Is Science compatible with Free Will?* New York: Springer. Chapter 7.

Hobson, J. A. (2005). Sleep is of the brain, by the brain and for the brain. *Nature, 437*, 1254–1256.

Kane, R. (2013). Can a traditional incompatibilist or libertarian free will be made consistent with modern science? Steps toward a positive answer. In A. Suarez & P. Adams (Eds.), *Is Science compatible with Free Will?* New York: Springer. Chapter 17.

Kant, I. (1785). Grundlegung zur Metaphysik der Sitten, 3. Abschnitt. http://gutenberg.spiegel.de/buch/3510/1; Groundwork for the Metaphysic of Morals, Chapter 3 http://www.earlymoderntexts.com/f_kant.html. Cited 19 September 2012.

Keller, F., & Iverson, J. M. (2013). The role of inhibitory control of reflex mechanisms in voluntary behavior. In A. Suarez & P. Adams (Eds.), *Is Science compatible with Free Will?* New York: Springer. Chapter 8.

Kent, A. (2010). One world versus many: The inadequacy of Everettian accounts of evolution, probability, and scientific confirmation. In S. Saunders, J. Barrett, A. Kent, & D. Wallace (Eds.), *Many worlds? Everett, quantum theory and reality*. New York: Oxford University Press. Chapter 10. http://arxiv.org/abs/0905.0624

Koch, C. (2009). A theory of consciousness. Is complexity the secret to sentience, to a panpsychic view of consciousness? *Scientific American Mind*, 16–19.

Libet, B. (2002). The timing of mental events: Libet's experimental findings and their implications. *Consciousness and Cognition, 11*(2), 291–299.

Lewis, C. S. (1947). Miracles: A preliminary study. Bless, G. London; reprint (1996) Harper Collins, New York. Chap. 3. www.harpercollins.com/browseinside/index.aspx?isbn13=9780060653019. Cited 22 September 2012.

Mele, A. (2013). Free will and neuroscience: Revisiting Libet's studies. In A. Suarez & P. Adams (Eds.), *Is Science compatible with Free Will?* New York: Springer. Chapter 13.

Merali, Z. (2013). Are humans the only free agents in the universe? In A. Suarez & P. Adams (Eds.), *Is Science compatible with Free Will?* New York: Springer. Chapter 6.

Popescu, S. (2012). Discussion at the Seminar: "The Greatest Inspiration Surely Is Nonlocality". Vall d'Illiez/Switzerland, June 1 (cited according to private communication of Do 21.06.2012 21:22).

Rizzolatti, G., & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: Interpretations and misinterpretations. *Nature Reviews Neuroscience, 11*, 264–274.

Staune, J. (2013). Towards a non-materialist realism. In A. Suarez & P. Adams (Eds.), *Is Science compatible with Free Will?* New York: Springer. Chapter 14.

Suarez, A. (2007). Classical demons and quantum angels: On 't Hooft's Deterministic Quantum Mechanics. http://arxiv.org/abs/0705.3974.

Suarez, A. (2011). Is this cell entity a human being? Neural activity, spiritual soul, and the status of the inner cell mass and pluripotent stem cells. In A. Suarez & J. Huarte (Eds.), *Is this cell a human being?* (pp. 171–192). Berlin: Springer.

Suarez, A. (2013). Free will and nonlocality at detection as basic principles of quantum physics. In A. Suarez & P. Adams (Eds.), *Is Science compatible with Free Will?* New York: Springer. Chapter 5.

Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *The Biological Bulletin, 215*, 216–242.

Tononi, G. (2013). On the irreducibility of consciousness and its relevance to free will. In A. Suarez & P. Adams (Eds.), *Is Science compatible with Free Will?* New York: Springer. Chapter 11.

Zeilinger, A. (2005). Interview. In: Die Weltwoche, Ausgabe 48/05, English at: http://www.signandsight.com/features/614.html.

Glossary

Authorship A description of the relation a human person has to a work or deed she brings about. Authorship presumes "personal identity" (see this Entry below) (Chaps. 5 and 7).

Autism Autistic spectrum disorder (ASD) is a syndrome characterized by impairment in social skills, communicative abilities, emotional responses, and motor behavior. Children with ASD have a severe impairment in motor organization that includes a deficit in chaining motor acts into intentional actions and, as a consequence, a lack of activation of intentional motor chains during action observation. ASD children, in order to understand the actions of others, do not use their internal motor knowledge, but another cognitive strategy. This evidence suggests that the mirror mechanism (see the Entry "mirror neurons" below) plays a strong role in mediating the capacity to understand the behavior of others and to entertain inter-individual interactions (Chap. 9).

Awareness The state of knowing one's own existence and the existence of other persons or entities (Chap. 10). For a perfect unlimited being there is no distinction between the act of being and the act of being aware. A fundamental limitation of a human being is that he/she cannot be aware or conscious all the time.

Backwards causation The hypothesis that it is possible to change the past by decisions in the present. So for instance in a Mach-Zehnder interferometer the hypothesis means that by choosing to change the path length *after* the particle leaves the first beam-splitter BS1, the experimenter can determine the path by which the particle leaves BS1 (Chap. 5).

Bell's inequalities Mathematical criterion of locality discovered by John S. Bell (in 1964–1965). The violation of Bell's inequalities by experiment means that there are phenomena in nature that cannot be explained by local causality, i.e., by signals traveling with a velocity less/equal to the velocity of light

---

The concepts contained in this Glossary mostly correspond to the keywords provided by the chapter authors. When suitable we refer to the chapter where the corresponding definition appears

A. Suarez and P. Adams (eds.), *Is Science Compatible with Free Will?: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience*,
DOI 10.1007/978-1-4614-5212-6,

(provided one excludes explanations of the "many worlds" type—see this Entry below) (Chaps. 2, 3, 4, 5 and 18).

Brain activity The activity produced by the neurons and neuronal networks responsible for determined motor or cognitive achievements. The centers responsible for spontaneous motility (see the Entry "Spontaneous movements") are located in the brain stem.

Brain dead The state of a human organism fulfilling the so-called "criteria of brain death" (mainly absence of spontaneous breathing, ocular and facial motor responses); on the basis of these practical recipes one establishes death in the clinical praxis. In a general theoretical way death is also defined as the "irreversible" breakdown of all brain activity including the brain stem. Nonetheless, a brain dead body may in fact maintain integrative capacities mediated by the brain to a certain degree. Therefore authors suggest that death means the "*irreversible* loss of the integrative capacity for spontaneous motility (animal behavior)", and this is what the standard clinical criteria of brain death actually attempt to ascertain (see Suarez 2011, referred to in Chap. 18). Even if physicians are not more able than physicists to explain what "irreversible" means precisely, the use of this adjective "irreversible" in the definition of death suggests damage that is beyond the human capacity to repair. A similar limitation seems to characterize also quantum measurement: once the collapse of the wave function occurs it is supposed that the experimenter cannot restore the initial state. For the time being nobody knows where this limitation comes from (measurement problem) (Chap. 5).

Causality The process by which a being brings another being into existence. So for instance when I hold a talk or edit a text I am the cause the words I speak or write. Regarding "causality" there is often a misunderstanding. In science "causality" is usually understood in the sense of causality within space-time or relativistic local causality: One says that event A is the cause of event B, if A precedes B in time (A lies in the light-cone of B), and the occurrence of A implies the occurrence of B. Accordingly, when one states that "quantum correlations cannot be causally explained," one means that the "correlations" cannot be explained by means of stories, i.e., information stored in space-time. However, such an expression does not mean that "quantum correlations don't have any explanation" or they "come from nothing." Indeed "correlations cry out for explanation" (John Bell), and in the case of the quantum ones the explanation lies outside space-time, that is, quantum correlations originate from a nonmaterial principle or agent. In other words, quantum physics denies causality in the sense of "temporal causality" (as the philosophers Hume and Kant understood it), but not in the sense of "agency from outside space-time." Therefore, it is misleading to state that quantum physics allow us to prove the universe originates from nothing.

Church of the larger Hilbert space Term coined by John A. Smolin to designate the community of scientists who believe that every event considered irreversible in a system is nothing other than a reversible or unitary evolution of a larger

system (Chap. 4). The Church is motivated by the wish to avoid the idea of the quantum collapse of the wave function proposed by the Copenhagen interpretation (Bohr, Heisenberg, Born, Jordan). In this sense one can say that the Church of the larger Hilbert space was very much inspired by Louis de Broglie's model of the local empty wave, subsequently developed by David Bohm as a model of the nonlocal quantum potential. The Church comes into existence with the "Many worlds interpretation" of Hugh Everett, according to which all possible outcomes of an experiment actually happen although in different worlds. John Bell felt that the "many worlds picture" is not only about the collapse but has something distinctive to say regarding nonlocality. This has been worked out by Lev Vaidman and more recently Gilles Brassard and Paul Raymond-Robichaud, who propose versions of many worlds assuming locality and determinism (Chap. 4). If the motivation of de Broglie's empty wave is to escape nonlocality in single-particle interference experiments (involving only two detectors), that of "parallel lives" (the many worlds version of Brassard and Raymond-Robichaud) is to escape nonlocality in two-particles entanglement experiments (involving four or more detectors). In a sense many worlds and parallel lives consequently develop the idea of the empty wave. And both, empty wave and parallel lives, rely on the assumption that there are entities existing and propagating within space-time which are inaccessible to observation. This is the distinctive "article of faith" (axiom) of the Church of the larger Hilbert space (Chap. 5).

Compatibilism The view that determinism in nature, and in particular in brain activity, is compatible with free will (Chaps. 4, 17 and 18). Compatibilism implies strong dualism, i.e. the complete separation of the philosophical, moral realm of freedom from the physical (neurophysiological) realm of nature.

Consciousness The state in which a person perceives or is aware of her own existence. Observation is basic for science. But observation cannot be defined without referring to an observer, and more precisely to the capacity of the observer of perceiving (becoming aware of) the data resulting from experiments. When does an outcome registration happen in an experiment? For the time being nobody can answer this question. It is the so-called "measurement problem" and likely to be the greatest challenge we are faced with in quantum physics. The "measurement problem" is often misunderstood as the necessity that there is no result (no "collapse of the wave-function") so long as an observer does not watch the apparatus (Schrödinger cat paradox). Actually, quantum physics does not state such a thing but only that to have a result an *irreversible* process has to take place in the apparatus, and this *irreversibility* is defined by conditions (a new constant of nature, or some mathematical condition) that make it possible for the result to be accessible to the human observer. Consciousness is not something one can explain by other things but something which is necessary to explain everything. An important point is that the human observer is not pure consciousness but a limited one; he/she is not always in a state of consciousness but only periodically. So consciousness is well defined by stating that it is what we lose every night when we fall asleep (see Chaps. 11 and 18). A being who is always

aware of his existence is surely more perfect than a being who realizes that he exists only from time to time. Limitation in consciousness is a sign of imperfection in being. The idea that *consciousness* first emerges through the evolution of matter (within space-time) when a certain degree of complexity is reached, overlooks the fact that observation is the foundation of science and has to be an ingredient of any attempt to explain the beginning of the universe scientifically.

Control, bottom-up Reflex control of eye movements. It describes attentional processing which is driven by the properties of the objects themselves. Reflex movements can be considered the effect produced automatically by an observable cause, the signal from salient external stimuli. Some events, such as a moving stimulus in visual field, or a sudden loud noise, can attract our visual attention in a pre-conscious or non-volitional way: we attend to them whether we want to or not (Chap. 8).

Control, top-down Volitional control of eye movements, also known as goal-driven, voluntary attention. Voluntary attention is under the control of the person and focused volitionally by signals derived from task demands like the instructions given by the experimenter to the proband (in addition to bottom-up mechanisms) (Chap. 8). Top-down control of eye movements can be considered a particular case of "top-down causation." This term is sometimes used to stress that certain behaviors cannot be explained "materialistically" by cause-effect reflex mechanisms and require control coming from a "higher" level than the pure physical or neurophysiological one. However this idea remains ambiguous and misunderstood, if one does not clarify that in this context the term "causation" is not used in the sense of the conventional deterministic causation. Therefore it may be convenient to stress that "top-down causation" to some extent involves "causation from outside space-time." And it is important to note as well that already the behavior of quantum mechanical devices in the lab cannot be explained "materialistically" by causal chains in space-time.

Determinism The hypothesis that each event A can be completely explained by the chain of events preceding A in time. Relativistic or local determinism restricts causality to causes lying in the light-cone of the effect under consideration.

Distal decision and intentions Decisions and intentions about things to do later (see the Entry "Proximal decisions") (Chap. 13).

Empty wave A notion introduced by Louis de Broglie to explain the quantum phenomenon of interference avoiding the Copenhagen or standard interpretation involving the collapse of the wave function. After any beam-splitter it is supposed that the particle (carrying energy and momentum) leaves by one of the two output ports, and an empty wave (without energy and momentum) leaves by the other output port. The empty wave records information about the path it travels and uses it to pilot the particle if they meet again in a second beam-splitter. So the model can account for interference assuming that the decision of the outcome happens at the beam-splitter, and escapes the conclusion of nonlocality implied by the idea that the decision about which detector clicks (the outcome) happens at detection. Nevertheless in experiments involving two

or more entangled particles the local empty wave cannot account for the quantum correlations, as Bell's theorem proves. It is important to note that the empty wave interacts only with "its particle" and is therefore undetectable. Thereby there enters into physics for the first time the idea of entities exiting and propagating in space-time that cannot be directly accessed by observation. In this context Einstein coined the term "ghost fields." De Broglie's "empty wave" developed later into Everett's "many worlds" (see this Entry below).

Entanglement The quantum state by which two (or more) physically separated systems must sometimes be thought of as a single (nonlocal) entity. A well-known example is the so-called *singlet* state in which two particles produce correlated detection outcomes even if they are space-like separated from each other, and the correlations cannot be explained by signals travelling at the velocity of light (Chaps. 2, 3 and 5).

Eye movements See the Entries: "control, bottom-up," "control, top-down," and "spontaneous movements."

First-person knowledge The capacity of knowing the intentions, wishes, feelings, emotions, of other persons by observing that they behave the same way as I do, that is, they exhibit bodily movements similar to those I perform to express my intentions, wishes, feelings, emotions (Chap. 9). Although a person has direct access only to her spiritual or mental powers (free will, consciousness), first-person knowledge indicates that a person by knowing herself also achieves knowledge about other persons.

Free will The power or capacity of a person to perform an action or make a choice without being completely determined either by a causal chain of events or information stored in space-time, or by other agents or entities existing outside space-time. For a human being to be free requires that his/her brain activity is not completely determined by past events. On the other hand free will requires that the outcomes of the brain activity do not produce purposeless and purely arbitrary behavior. Hence a certain degree of indeterminism (quantum randomness) at the level of brain activity is a necessary but not sufficient condition for human free will. However, quantum randomness does not exclude control, on the contrary it is always accompanied by a certain degree of control coming from outside space-time. Thus quantum physics provides a description of the world that fits well with the assumption of free will. What is more, quantum physics itself assumes the free will of the experimenter as an axiom. In this sense free will is not a principle that can be explained by other things, but a principle which is necessary to explain everything. The position that "free will is an illusion" is itself an illusion arising from the fact that human capacity for purposeful behavior is limited (not abolished) by a number of factors, in particular by the need to sleep. In summary, quantum physics allows us to overcome the deterministic objection to free will without reducing free will to arbitrariness. Obviously quantum physics does not solve the classic theological problem about the compatibility of human free will with the omniscience of God.

Human being An individual of the human species under the aspect that he or she is the subject of rights, especially the right to life and property, and deserves the corresponding respect on the part of other human beings. In this sense, human being can be considered synonymous with "human person."

Human person Often understood as an "individual substance of a rational nature" (definition according to Boethius). From the perspective of observable operations or signs allowing us to determine whether a living organism is a human person, the definition of person is linked to the capability for developing corporal operations similar to those one individual of the human species uses to communicate with other individuals of the human species. This position defines "person" through "relation," and assumes it is a basic category for understanding and explaining the world; that is, one defines animals as living beings exhibiting movements like human spontaneous ones (Chap. 18 and References therein).

Human soul The form that is proper to the human body (as defined according to Aristotle and Thomas Aquinas), viewed as the immaterial integrating principle of a living human organism capable of free will and knowledge. The presence of the human soul is *directly* revealed by the integrated and coordinated operations proper to an organism with spontaneous motility (animal behavior), and it can likewise be deduced from the observable biological features ensuring the capability to develop spontaneous movements. Even if the human soul as such cannot be directly accessed by experimental procedures, the body of the human species with spontaneous motility is a visible sign of the soul; it is nothing other than the embodied presence of the soul in space and time. By contrast a brain dead human, even if it exhibits a certain degree of integrated functioning, still lacks the proper biological potential for performing spontaneous movements and therefore it does not share the moral status of a person (Chap. 18 and References therein).

Indeterminism The hypothesis that the outcome of a quantum experiment cannot be completely explained (even in principle) by the chain of events preceding the outcome in time. Quantum indeterminism is intrinsically united to nonlocality and agency from outside of space-time (Chaps. 2 and 3).

Incompatibilism The position stating that determinism in nature (and therefore in the brain) excludes the possibility of human free will and freedom. Incompatibilists consider indeterminism in nature a necessary condition for human free will without reducing free will to indeterminism (Chaps. 4, 17 and 18).

Inhibition See voluntary inhibition.

Integrated Information Theory (IIT) A theory of consciousness proposed by the psychiatrist and neuroscientist Giulio Tononi. According to IIT, when a choice is made consciously, the choice is maximally irreducible and cannot be attributed to anything less than the entire complex (for example, a network of neurons, some firing and some not) that brings it about. A conscious choice, while maximally and irreducibly causal, is also necessarily under-determined and thus unpredictable. According to IIT, indeterminism is not to be thought of as "a sprinkle of randomness that instills some arbitrariness into a preordained cascade

of mechanisms, decreasing their causative powers." Rather, indeterminism provides a backdrop on which the entire complex acts to impose self-control, understanding and alternative possibilities and generates integrated information (Chap. 11).

Libet's experiments Experiments originally performed by Benjamin Libet, who interpreted the results as a demonstration that the brain activity (readiness potential) responsible for a motor action precedes the conscious decision of the subject to perform this action. One main result of this book is that these experiments neither obliterate free will nor responsibility, but rather prove that human free will and consciousness are limited (Chaps. 11, 12, 13, 14 and 15).

Locked-in syndrome The condition in which a patient is conscious but cannot communicate with others because of muscular lesions that make him or her incapable of performing spontaneous movements (total locked-in). Patients with partial locked-in syndrome can communicate using eye movements. Distinguishing total locked-in syndrome from persistent vegetative state is a challenging objective of intensive current research (Chap. 10).

Many worlds An interpretation of quantum mechanics proposed by Hugh Everett, who postulates that all possible outcomes of an experiment actually happen although in different worlds. When a particle enters a beam-splitter there is not only the outcome of a particle leaving by output port A and an empty wave by output port B, but also the alternative outcome of a particle leaving by output port B and an empty wave by output port A. However, the latter outcome happens in a different world: at any choice device the world splits into as many worlds as available choices, and after the split these worlds remain inaccessible to each other. Many worlds accounts for a unitary evolution of the quantum state and even for the quantum correlations invoking exclusively local causality, i.e. causal links within the light-cone. However, the postulate of the "world split" implicitly assumes the idea of two resulting space-time manifolds which remain inaccessible to each other even if they cannot be considered space-like separated. This view is further developed by the "parallel lives" interpretation (see the corresponding Entry below) (Chaps. 3, 4 and 5).

Mirror neurons Nerve cells discovered in the premotor and parietal cortex, which become active during observation and execution of motor acts. The mirror system unifies action production and action observation, and is the neurophysiological basis for the understanding of another's actions and intentions from the inside (first-person knowledge). Mirror neurons provide a neurophysiological correlate of the spontaneous movements' criterion for ascertaining personhood (Chaps. 9 and 18).

Neuroeconomics Neuroeconomics is a combination of psychology, economics, and neuroscience, which aims to understand, at the most basic level, how individuals make economics decisions. The idea is to measure and record brain activity when an individual must choose between various economic options. Neuroeconomics provides a "platform" for a theory of deterministic human behavior: the idea that if we are able to measure brain activity well enough, then economic behavior will be predictable (Chap. 15).

Neuron A nerve cell; an excitable cell specialized for the transmission of electrical signals (nerve impulses) over long distances within the body.

Nonlocality The quantum notion used to describe phenomena that cannot be explained by relativistic local causality, that is, by signals propagating at the velocity of light, or slower. In the Solvay conference (1927) Einstein objected to the idea of the collapse of the wave function arguing that it implies action at a distance and conflicts with relativity. Einstein used a single-particle gedanken-experiment demonstrating that by the quantum collapse one is led to nonlocality at detection. An important aspect Einstein did not explicitly mention is that the experiment also shows that without nonlocality one would violate the conservation of energy in each single quantum event. In addition, de Broglie's idea of the empty wave allowed the explaination of single-particle interference without invoking nonlocality. After 1927 Einstein withdrew from this argument and joined the EPR one using a two-particle entangled state. With EPR Einstein could contest quantum nonlocality without contesting the conservation of energy. The EPR argument was further developed by David Bohm (1952) and led finally to the discovery of locality criteria by John Bell, the so-called Bell's inequalities. Experimental violation of these inequalities allows us to decide between Einstein's local view and quantum nonlocality in two-particle experiments even under the assumption of empty waves. That is, experiments violating Bell's inequalities refute the explanation of quantum effects by means of local inaccessible empty waves, referred to as local hidden variables. In addition, violation of Bell's inequalities does not relate to conservation of energy, and cannot be implemented to prove nonlocality in single-particle experiments (using only two detectors). Hence nonlocality at detection seems to be more basic than Bell's nonlocality. Anyway, the question of whether they are two different types of nonlocality or that one of them derives from the other is an open question (Chaps. 2, 3, 4 and 5).

Parallel lives A version of the "many worlds interpretation" of quantum mechanics proposed by Gilles Brassard and Paul Raymond-Robichaud (Chap. 4). These authors consider that the domains resulting from the "split" of the world at the choice device can be considered parallel regions within the same space-time. Thereby "parallel lives" objects mainly to nonlocality (instead collapse): Violation of Bell's inequalities means refutation of local hidden variables of the empty wave type, but it does not prove nonlocality coming from outside space-time. "Parallel lives" is based on the assumption that entities propagating within space-time can in principle be inaccessible to observation, an assumption implied by "empty wave" (Chap. 18).

Persistent vegetative state The state of patients exhibiting spontaneous movements but (in contrast to locked-in patients) lacking consciousness and incapable of spontaneous movements to communicate with others.

Personal identity A description of the way the human person exists in time: the person does not change, even though his or her body and personality can develop in time. In addition, by assuming personal identity and authorship (crucial for granting personal rights) one is assuming agency coming from outside

space-time: You, and the paper you are writing now, cannot be explained exclusively by material or observable causal chains (Chap. 5).

Proximal decisions and intentions Decisions and intentions about things to do straightaway. Ann's decision to phone Al now is a proximal decision; her decision to phone Bob tomorrow is a distal decision. Deciding to do something should be distinguished from wanting (or having an urge) to do it (sometimes people want to do things that they decide not to do). In the context of Libet's experiments decisions and intentions are supposed to be proximal ones. One may question Libet's assumption that the urge to perform a hand flex is an unconscious proximal decision to flex, or the readiness potential is a proximal cause of the flex instead of a (necessary but not sufficient) condition for it (Chap. 13).

Quantum nonmaterial agency Quantum effects, as described in Chaps. 3, 5, and 18, reveal nonmaterial agency: quantum interference and correlations come from outside space-time in the sense that they cannot be explained exclusively by stories that are stored in space-time, that is by material agency.

Randomness The term is often used in the sense of something that happens by chance (without cause), and lacks any order or control. In quantum physics randomness means rather the quality of phenomena that cannot be explained by causal links in space-time. So for instance in entanglement experiments the outcomes on Alice's side exhibit a uniform random distribution like the tosses of a fair coin, and the same holds for the outcomes on Bob's side. However, Alice's outcomes are correlated with Bob's ones: randomness here and randomness there, but the same randomness at both sides. Quantum local randomness is inseparably united to nonlocal co-ordination. To generate true randomness one needs free will, and production of quantum randomness happens through expansion of free will (Chaps. 2 and 3), and this suggests that quantum randomness can be considered a particular case of free will (Chap. 5).

Realism The view that the quantum state or the quantum wave function reduces to the content of a scientist's knowledge, that is, it does not exist outside the scientist's mind. Realism is used (often equivocally) in two different senses: in a *materialistic* or *deterministic* one, if one assumes that the quantum state is an entity existing in space-time; or an *ontic* one, if one assumes that "not all that is important for physical phenomena is contained in space-time." *Materialistic* realism is at odds with standard quantum physics and in the end is equivalent to superdeterminism, i.e., the view that all that happens, even the experimenter's decisions, are predetermined since the origin of the universe. The sense in which Bernard d'Espagnat refers to the quantum reality as "veiled reality" (réel voilé) seems near to *ontic* realism.

Reflex movements The movements of the body resulting automatically as the effect of an observable stimulus (see Entry "bottom-up control").

Responsibility (ultimate) The idea that an agent is responsible for anything that is a sufficient reason (condition, cause, or motive) for the occurrence of an action. If the agent is at least in part responsible by virtue of choices or actions voluntarily and consciously performed in the past for having the character and

motives he or she now has when performing a choice, then the agent can be considered to be ultimately responsible for the choice. Ultimate responsibility highlights the fact that the free will issue is about the freedom of the will and not merely about the freedom of action (Chap. 17).

Rights A distinctive human feature. Rights immediately refer to their own body. By claiming rights I primarily want others to respect my body, and this implies that I myself have also to respect any body that belongs to the same species as me. The sense of rights is crucial to understand and explain the world. You cannot separate what you are and know from the wish to be respected by others and the consequent duty to respect others. A science or a metaphysics that does not allow you to defend your rights remains a useless piece of intellectual work. Although Darwinian rules play a role in bringing about the human species, according to Richard Dawkins this species ought to organize its intraspecific life according to "anti-Darwinian" principles [http://www.abc.net.au/tv/qanda/txt/s3469101.htm]. Does this mean that the "sense of rights" and "responsibility" (see this Entry) are principles that do not arise (at least entirely) through Darwinian evolution but rather against it? If the answer is yes this could mean that these principles (like free will and consciousness) come, in the end, from outside space-time.

Self-forming actions Acts in our past life histories by which we formed our present characters and with respect to them we could have done otherwise than we did (Chap. 17).

Spontaneous movements Movements of the human body similar to the movements one human person makes for expressing thoughts, emotions, wishes, and rights-claims to other human persons (Chap. 18). Even if they are often unconscious and unintentional, spontaneous movements are potentially will-directed movements, that is, they are movements that can always be directed by the will when chosen (Suarez 2011 in Chap. 18). The conventional classification as reflexes, automatic and voluntary (conscious) movements does not cover a number of motor behaviors like the rapid eye movements occurring in REM sleep, facial and leg movements exhibited by persistent vegetative state (PVS) patients or those with Huntington's chorea. Since they are considered "involuntary" (because they are unconscious) sometimes they are unfittingly described as "reflexive responses to internal stimuli," since by "internal" one means stimuli that are not accessible to observation. Therefore, we think one should distinguish between *conscious* and *unconscious* voluntary movements. *Unconscious* voluntary movements we denote *spontaneous* as well: Like the conscious ones, spontaneous movements are in principle unpredictable (by others) and reveal nonmaterial agency (from outside space-time).

Voluntary inhibition The capacity of forming a behavior or habit by inhibiting reflexes arising from external stimuli or spontaneous movements resulting from internal urges corresponding for instance to archaic evolutionary patterns of behavior.

Voluntary movements Usually understood in the sense of conscious free-willed movements. However, there are unconscious spontaneous movements (see this Entry above) that are neither reflexes nor autonomous motor acts. Such spontaneous movements may well be denoted as voluntary unconscious movements.

Veto The hypothesis that in Libet's experiments the subjects can inhibit an action (a wrist flex assumed to occur at a time labeled 0 s) *after* they become conscious (at −0.2 s) of the urge to act (initiated by the readiness potential at −0.55 s). This means the subjects have a window of about 150 miliseconds to impose their free will against the unconscious urge to flex. According to Libet there is a conscious veto of the action (*after* −0.2 s) when there is no flex, and there is an unconscious proximal cause (the readiness potential at −0.55 s) and no conscious proximal decision at all when there is a flex. By contrast, according to the view of "voluntary inhibition" there is a conscious proximal decision of "no inhibition" (at −0.2 s) when there is flex, and there is a conscious proximal decision to inhibit (at −0.2 s) when there is no flex (Chaps. 8, 13, 14).

Wake–sleep cycle A fundamental feature of nature that makes it possible to have limited consciousness. The Ascending Reticular Activating System (ARAS) located in the brain stem plays a key role in ensuring the wake–sleep cycle. Being awake enables self-control to produce behavior adapted to the sense data coming from the external environment. The parameters or constants of nature responsible for the wake–sleep cycle are to date unknown.

Index

A. Suarez and P. Adams (eds.), *Is Science Compatible with Free Will?: Exploring Free Will and Consciousness in the Light of Quantum Physics and Neuroscience*,
DOI 10.1007/978-1-4614-5212-6, © Springer Science+Business Media, LLC 2013